Food Search Based on User Feedback to Assist Image-based Food Recording Systems

Sosuke Amano
Dept. of Information and Communication Eng. The University of Tokyo
7–3–1 Hongo, Bunkyo–ku
Tokyo 113–0033, JAPAN
s_amano@hal.t.u-tokyo.ac.jp

Shota Horiguchi
Dept. of Information and Communication Eng. The University of Tokyo
7–3–1 Hongo, Bunkyo–ku
Tokyo 113–0033, JAPAN
horiguchi@hal.t.u-tokyo.ac.jp

Kiyoharu Aizawa
Dept. of Information and Communication Eng. The University of Tokyo
7–3–1 Hongo, Bunkyo–ku
Tokyo 113–0033, JAPAN
aizawa@hal.t.u-tokyo.ac.jp

Kazuki Maeda
foo.log Inc.
Entrepreneur Plaza 701,
7–3–1 Hongo, Bunkyo–ku
Tokyo 113–0033, JAPAN
maeda@foo–log.co.jp

Masanori Kubota
foo.log Inc.
Entrepreneur Plaza 701,
7–3–1 Hongo, Bunkyo–ku
Tokyo 113–0033, JAPAN
kubota@foo–log.co.jp

Makoto Ogawa
foo.log Inc.
Entrepreneur Plaza 701,
7–3–1 Hongo, Bunkyo–ku
Tokyo 113–0033, JAPAN
ogawa@foo–log.co.jp

ABSTRACT

Food diaries or diet journals are thought to be effective for improving the dietary lives of users. One important challenge in this field involves assisting users in recording their daily food intake. In recent years, food image recognition has attracted a considerable amount of research interest as a new technology to help record users' food intake. However, since there are so many types of food, and it is unrealistic to expect a system to recognize all foods. In this paper, we propose an optimal combination of image recognition and interactive search in order to record users ' intake of food. The image recognition generates a list of candidate names for a given food picture. The user chooses the closest name to the meal, which triggers an associative food search based on food contents, such as ingredients. We show the proposed system is efficient to assist users maintain food journals.

Keywords

food journal; image recognition; associative search;

1. INTRODUCTION

Food diaries or dieting journals are considered an effective approach to improving one ' s dietary life. Various services are available nowadays to assist users with recording and analyzing their daily food intake. In recent years, smartphones have played an important role as a portable food recording tool. Services, such as MyFitnessPal [3] and Asken [1], have developed smartphone applications that help users record their food intake.

Food journaling requires that users record each meals, along with the time and amount of consumption. The discharge of this operation imposes stringent demands on the user. To resolve this problem, past studies have proposed a number of image-based food recording systems [8, 13, 18, 5, 9]. These systems assume the following: A user is asked to take a picture of the food. The system then suggests names for the meal through image recognition. Finally, the user records the meal name by selecting the correct one in the list of suggestions. Although these system recognize a limited number of meals at a high accuracy, it remains difficult to recognize meals accurately enough to serve the purposes of food journaling. Image recognition requires a large number of pictures to train its models. However, the variety of meals in everyday life is far too large for there to be an exhaustive database of images to train learning models.

In this paper, we propose a method to update the result of image recognition by associative search based on user feedback. An overview of the proposed system is shown in Figure 1. The system first suggests meal names by recognizing a photo of the food taken by the user. The user then selects the most appropriate name. If the list does not contain the correct name of the given meal, the user can select the closest one. Following this, the system updates the list by searching associative meals in the entire database. By a combination of food image recognition and food search, the user can thus search a much larger number of meals beyond the classes that can be recognized from a picture.

2. RELATED WORK

Food image recognition is considered a core task for image-based food recording systems [9, 25, 7, 11, 13, 8]. Recent successful approaches [25, 8] are based on Convolutional Neural Networks (CNN), such as AlexNet [14] and GoogLeNet [21]. These recognition models are trained by such food image datasets as Food—101 [9] and UEC–FOOD256 [12].

MADiMa'16, October 16 2016, Amsterdam, Netherlands

© 2016 ACM. ISBN 978-1-4503-4520-0/16/10. . . $15.00

DOI: http://dx.doi.org/10.1145/2986035.2986037

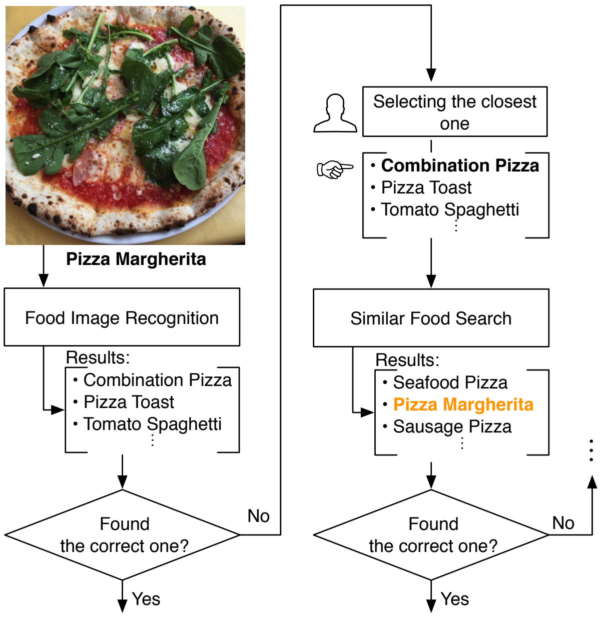

Figure 1: An overview of the proposed system. It performs image recognition on a given picture of a food and shows a list of candidates. The user finds the appropriate one in the list. If not, he/she selects the most associated one with the given dish. The system then updates candidates by associative search of foods.

However, these datasets are not large enough to accommodate all meals featuring in food journaling applications. Of food nutrition databases, Eat Smart [2] contains 1,870 classes of general foods in Japan and more than 90,000 in their detailed database. The USDA ’ s National Nutrient Database [22] contains 8,789 classes of standard food from the United States. As a purportedly exhaustive food dataset, MyFitnessPal [3] provides users with millions of foods, along with their nutritional and caloric information. It is impractical to recognize these large numbers of classes using only images.

In this paper, we propose a method to improve the results of image recognition of food by associative search based on user feedback. It is similar to Relevance Feedback (RF) [24, 20, 17], which is a classical approach to the problem whereby users' search queries are often insufficient to understand their intentions in information retrieval systems. It attempts to complete these queries by analyzing users' actions on initial search results. Various retrieval targets and user feedback types have been discussed in the literature. Rui et al. [20] applied RF to content-based image retrieval. They defined five levels of relevance for each item. The function representing the affinity between two images is updated by a relevance score obtained from the initial images. In our proposal, we define a relevance based on food content, such as ingredients and nutritional value.

Our RF searches for foods associated with the image of a given meal from an entire text database of food. Computing similarity between two meals has been discussed in the field of food recipe analysis. Some previous work has reported that ingredients and descriptions of recipes are more important than the names of meals to search similar recipes [23, 10]. Therefore, we consider this information in the food database.

3. DATASET

3.1 Eat Smart database

We used the Eat Smart database provided by Eat Smart Inc. [2] as our generic food nutrition database. This database contained data regarding two types of food: a generic database containing 1,870 common meals, such as “egg sandwich,” and a very detailed food database containing approximately 90,000 meals, such as “egg sandwich sold at 7-eleven.” We used only the generic database in this study.

The database includes information concerning the ingredients of foods, and the amount of the various ingredients and nutrients in each meal. Almost all ingredients were defined by the Standard Tables of Food Composition in Japan [19].

Each meal was assigned to a major and a minor category. There were 17 classes in the major category and 91 in the minor. For example, “custard pudding” was categorized in “sweets and dessert” as its major category, and “cake, jelly, and parfait” as its minor category.

3.2 FoodLog dataset

To construct the food image recognition system, we used the data from FoodLog [5]. FoodLog is an image-based food journaling application that was launched in July 2013. It is mainly used in Japan. FoodLog users record their daily food intake to the FoodLog system via photos. Meals in a photo are localized as square areas through the touching and dragging operation on a smartphone. Each area is annotated by a meal name. The meal names have two types of definitions. One is from the generic database of Eat Smart and the other from users' free description. We only considered the former in our recognition system.

4. OUR PROPOSED SYSTEM

An overview of our proposed system is provided in Figure 1. The system first applies CNN to recognize a meal in the image given by a user and suggests meal names from a list. The user chooses the correct meal name from the list. In case the meal is not listed, the user can select the meal most relevant with the given one. The system then searches for associative meals in the database to update the list of candidate meal names.

4.1 Food image recognition

In order to implement food image recognition, we made use of Network In Network (NIN) [16] which is a CNN with a small network structure consisting of inexpensive computational resources. In order to optimize the determination of what to recognize and how many classes to recognize, we compared several models in the experiments below.

4.2 Associative Food Search

The meals in the generic dataset are divided into categories, and the ingredients of each meal are available. We defined the degree of association by using categories and ingredients.

We compute text vectors and ingredient vectors. Text vector T_i of meal i is generated by its name, and an ingredient vector I_i is generated by its ingredients. Moreover, we define a category label c_i. The degree of association $a(i,j)$ between meal i and j is defined as:

$$a(i,j) = cos(T_i, T_j) + w_I \cdot cos(I_i, I_j) + w_c \cdot d(c_i, c_j) \quad (1)$$

where $cos(x,y)$ is cosine similarity between x and y. $d(x,y) = 1$ if $x = y$, and is -1 otherwise. w_I and w_c represent the importance of the ingredients and the categories, respectively. We chose $w_I = 1$ and $w_c = 0.5$ in the experiments below.

4.2.1 Text vector

We compute a bag-of-word representation of character trigrams as text vectors. Before computing a character trigram, meal names are transformed into *kana*, which indicates the pronunciation of the word in Japanese. To estimate *kana* from raw Japanese text, we used Mecab [15] with the Unidic [4] Japanese parts-of-speech dictionary.

4.2.2 Ingredient vector

The Standard Tables of Food Composition in Japan consists of six layers of food groups. For example, flour is defined as “Cereals, cereals, common wheat, flour, hard flour, first grade” in the table. The top layer has 18 food categories such as “Cereals”. Therefore, the vector representing the top layer has 18 dimensions. The six vectors, from the top to the bottom layers, are connected to form one vector.

To normalize the number of different food groups and layers, we compute the cumulative frequency $f_g(i)$ from the amount $x_g(i)$ for a meal i and a food group g. $x_g(i)$ is first normalized into $\hat{x}_g(i) = x_g(i)/w_g$, where w_g is defined as the average amount of the food group g.

$$w_g = \frac{\sum_{x_g(i)!=0} x_g(i)}{\sum_{x_g(i)!=0} 1} \quad (2)$$

Then, we compute the cumulative frequency $f_g(i)$:

$$f_g(i) = CDF_{\mu,\sigma^2}(\hat{x}_g(i)) \quad (3)$$

CDF_{μ,σ^2} is the cumulative density function of the normal distribution, with mean μ and variance σ^2. μ and σ^2 are computed by the distribution of $\hat{x}$. Finally, the g-th value of the ingredient vector $I(i)$ is defined as:

$$I_g(i) = f_g(i) - CDF_{\mu,\sigma^2}(0) \quad (4)$$

This value is 0 if the amount of food group g is 0. It changes from 0 to $1 - CDF_{\mu,\sigma^2}(0)$ with a sigmoid-like curve.

5. EVALUATION

5.1 Food image recognition

To optimize image recognition in the proposed system, we compared six models trained for different classes of meals. The classes used for the first three models consisted of 200, 400, or 800 most frequently eaten meals according to the Eat Smart database (R-200, R-400, and R-800), and those used for the other three models consisted of 200, 400, or 800 most frequently eaten meal groups generated by clustering meal names of real use (C-200, C-400, C-800). For example, both “miso soup with radish” and “miso soup with nameko mushroom” are considered “miso soup” in the C-200. To create the lookup table from meals to groups, we first applied the clustering method proposed by S. Amano [6], and then manually checked the lookup table by focusing on frequently invoked groups and modified errors in the clustering process. Note that this conversion was applied to the training, the testing, and the evaluation steps.

We prepared the training, the test, and the evaluation datasets. The first two consisted of FoodLog data between Aug. 1, 2013 and Aug. 1 2015. We randomly divided them into a 9:1 ratio for training and testing, respectively. The number of data items varied with model. We trained the CNN model and tested each model using these datasets. The evaluation dataset was chosen from 490,140 instances of meals items from the FoodLog dataset between Aug. 1, 2015 and Feb. 1 2016.

The evaluation results are shown in Table 3. We computed three scores: “coverage,” “closed test accuracy,” and “open test accuracy.” Coverage means the percentage of trained classes in the evaluation dataset. Closed test accuracy represents the accuracy of the top five items from the data within the trained classes. Open test accuracy means the top five most accurate items in the entire evaluation.

The coverage increased from R-200 to R-800, and from C-200 to C-800, due to the increasing the number of training classes. The coverage of C was larger than that of R due to clustering. On the contrary, closed test accuracy decreased against the coverage. Open test accuracy results indicated the combination of these two factors. In our experiments, either C-200 or C-400 was the most suitable class to train the image recognition model.

5.2 Associative Food Search

We evaluated the association function by four evaluators. We first randomly selected 50 meal names as food search queries from the Eat Smart database. To test the proposed method, we compared it against three baseline methods that computed the degree of association only using one of three factors: “text,” “ingredient,” or “category.” The methods based text or ingredient computed cosine similarity through their vectors. The category-based method randomly selected meals in the same category as the one used in the query. The method then suggested five candidate meals for four evaluators. All evaluators were Japanese males in their 20s. The evaluators selected meals that were not relevant to the queries. Note that relevance was defined with reference to subjective evaluation. The evaluators were allowed to collect information from Web search engines if they did not know the meals used in the experiments.

A few examples of search and evaluation results are shown in Table 1 and 2. There are some conflicts in the results. One reason for this is the ambiguity in the meaning of the term “association.” For example, two evaluators determined that “Orange Juice” was relevant to the target “Cassis and Orange,” and two others disagreed. In this case, both are beverages made from orange, but one of them is non-alcoholic.

The aggregated result is shown in Table 4. The proposed method outperformed the other three baseline methods for all evaluators. The result showed that the fusion of different types of data is required to compute association among meals. For example, Table 1 shows that only the proposed method indicated both the classes “orange” and “alcoholic beverage” for the query “Cassis and Orange”. The text- and ingredient-based methods only categorized it in the former, and the category-based method only considered the latter.

Table 1: Examples of 5 top candidates to the query "Cassis and Orange" by four associative food search methods.

	Proposed	Text	Ingredient	Category
1st.	Campari and Orange	Orange	Campari and Orange	Amazake
2nd.	Fuzzy Navel	Campari and Orange	Fuzzy Navel	Shochu (Otsu)
3rd.	Screw Driver	Orange Jelly	Orange Juice	Wine Cooler
4th.	Mimosa	Cassis and Soda	Orange Drink (30% juice)	Shochu (Ko)
5th.	Orange Fizz	Cassis and Oolong Tea	Orange Drink (50% juice)	Draft Beer

Table 2: Evaluations of the degree of association by four evaluators (A, B, C, and D).

Query	Answer	A	B	C	D
Cassis and Orange	Orange Juice	○	×	×	○
	Draft Beer	○	×	○	○
Squid with Tomato Sause	Grilled Squid with Sea Urchin Roe	○	○	○	×
	Simmered Skin-on Japanese Amberjack	×	×	○	○

Table 3: The top 5 accuracy values of different target classes with the same network structure (NIN).

Class	Training image (per class)	Coverage	Closed Test	Open Test
R-200	855	45%	0.698	0.312
R-400	495	56%	0.582	0.327
R-800	180	67%	0.377	0.251
C-200	990	57%	0.680	0.384
C-400	495	69%	0.563	**0.387**
C-800	135	78%	0.313	0.245

Table 4: The evaluation results for association functions. This comparisons shows the precision of 5 answers by four evaluators (A, B, C, and D).

Method	A	B	C	D
Proposed	**0.892**	**0.66**	**0.752**	**0.864**
Text	0.696	0.476	0.52	0.668
Ingredient	0.796	0.548	0.576	0.668
Category	0.636	0.384	0.544	0.616

This shows that the proposed method satisfactorily handled the ambiguity of association among foods.

6. APPLICATION

We demonstrate the proposed system in figure 2. We used the C-400 model trained at section 5.1. Since the model recognizes the clusters of meals, we picked up one representative meal for each cluster. Initial candidates are shown as the list of representatives of the results of image recognition.

From the top to bottom of the figure 2, the system was given input images of "Grill Shrimp in the shell", "Cold Dam Dam Noodle", "Phad Bai Gaprao" and "Minestrone". In the case of "Minestrone", the initial candidate via image recognition is correct. In the case of top row of figure 2, "Salt-grilled Shrimp" among the initial candidates is the most associated. Then, the user triggers associative search, which shows the correct "Grilled Shrimp in the Shell" among the updated candidates. The second row of figure 2 is another example in which the associative search works.

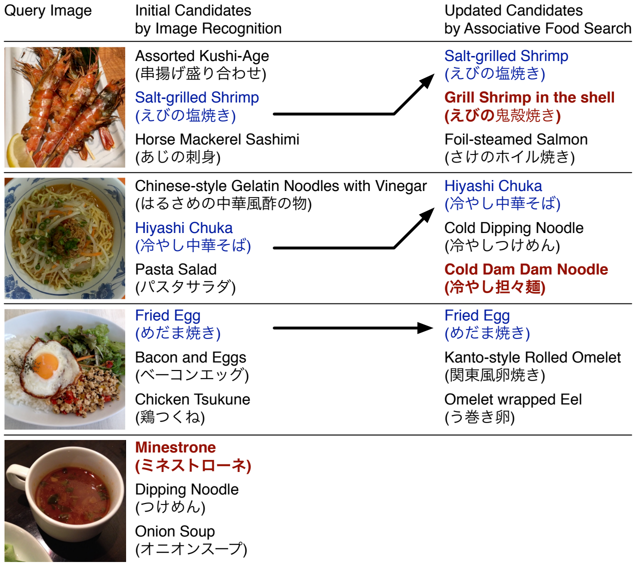
Query Image
Initial Candidates by Image Recognition
Updated Candidates by Associative Food Search
Assorted Kushi-Age (串揚げ盛り合わせ)
Salt-grilled Shrimp (えびの塩焼き)
Horse Mackerel Sashimi (あじの刺身)
Salt-grilled Shrimp (えびの塩焼き)
Grill Shrimp in the shell (えびの鬼殻焼き)
Foil-steamed Salmon (さけのホイル焼き)
Chinese-style Gelatin Noodles with Vinegar (はるさめの中華風酢の物)
Hiyashi Chuka (冷やし中華そば)
Pasta Salad (パスタサラダ)
Hiyashi Chuka (冷やし中華そば)
Cold Dipping Noodle (冷やしつけめん)
Cold Dam Dam Noodle (冷やし担々麺)
Fried Egg (めだま焼き)
Bacon and Eggs (ベーコンエッグ)
Chicken Tsukune (鶏つくね)
Fried Egg (めだま焼き)
Kanto-style Rolled Omelet (関東風卵焼き)
Omelet wrapped Eel (う巻き卵)
Minestrone (ミネストローネ)
Dipping Noodle (つけめん)
Onion Soup (オニオンスープ)

Figure 2: Recording meal names in our system. The meals highlighted in blue are user feedback that the user selects as the closest one among the initial results. The meals highlighted in red are the correct names.

The system fails in the case the initial candidates are very far from the correct one. The third row is the failure example. Its correct meal name is "Phad Bai Gaprao", but the initial candidates are not associated enough to find the correct one.

7. CONCLUSIONS

In this paper, we proposed the interactive search in order to assist image-based food recording systems. In case the correct meal name is not suggested by recognizing the picture of the meal, the user can select the closest name to the meal, which triggers the associative search using larger scale food database. We showed that the fusion of text, ingredient and category of meals is important to compute the degree of association between meals in a food database.

Although it is important to optimize the type and number of classes recognized as initial candidates, this has not been sufficiently been discussed in this work. Additional research is needed to construct better systems that use visual and text information simultaneously to assist users maintain food journals.

8. REFERENCES

[1] Asken diet. http://www.asken.jp/, July 2016.

[2] Eatsmart. https://www.eatsmart.jp/, July 2016.

[3] Myfitnesspal. https://www.myfitnesspal.com/, July 2016.

[4] Unidic. https://osdn.jp/projects/unidic/, July 2016.

[5] K. Aizawa, K. Maeda, M. Ogawa, Y. Sato, M. Kasamatsu, K. Waki, and H. Takimoto. 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*, 8(2):203–208, March 2014.

[6] S. Amano, K. Aizawa, and M. Ogawa. Food category representatives: extracting categories from meal names in food recordings and recipe data. In *Multimedia Big Data (BigMM), 2015 IEEE International Conference on*, pages 48–55. IEEE, 2015.

[7] S. Ao and C. X. Ling. Adapting new categories for food recognition with deep representation. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1196–1203, Nov 2015.

[8] M. Bolaños and P. Radeva. Simultaneous food localization and recognition. *arXiv:1604.07953*, 2016.

[9] L. Bossard, M. Guillaumin, and L. Van Gool. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014*, pages 446–461. Springer, 2014.

[10] S. Hanai, H. Nanba, and A. Nadamoto. Clustering for closely similar recipes to extract spam recipes in user-generated recipe sites. In *Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services*, page 31. ACM, 2015.

[11] H. Kagaya, K. Aizawa, and M. Ogawa. Food detection and recognition using convolutional neural network. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 1085–1088. ACM, 2014.

[12] Y. Kawano and K. Yanai. Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In *Computer Vision-ECCV 2014 Workshops*, pages 3–17. Springer, 2014.

[13] Y. Kawano and K. Yanai. Foodcam: A real-time food recognition system on a smartphone. *Multimedia Tools and Applications*, 74(14):5263–5287, 2015.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, pages 1097–1105, 2012.

[15] T. Kudo, K. Yamamoto, and Y. Matsumoto. Applying conditional random fields to japanese morphological analysis. 4:230–237, 2004.

[16] M. Lin, Q. Chen, and S. Yan. Network in network. *Proceedings of the International Conference on Learning Representations*, 2014.

[17] Y. Liu, T. Mei, X.-S. Hua, J. Tang, X. Wu, and S. Li. Learning to video search rerank via pseudo preference feedback. In *2008 IEEE International Conference on Multimedia and Expo*, pages 297–300. IEEE, 2008.

[18] A. Meyers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. P. Murphy. Im2calories: Towards an automated mobile vision food diary. December 2015.

[19] Office for Resources, Policy Division Science and Technology Policy Bureau, editor. *Standard Tables of Food Composition in Japan (2010)*. Official Gazette Co-Operation of Japan, 2010. (In Japanese).

[20] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Transactions on circuits and systems for video technology*, 8(5):644–655, 1998.

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[22] N. D. L. US Department of Agriculture, Agricultural Research Service. Usda national nutrient database for standard reference, release 28. http://www.ars.usda.gov/nea/bhnrc/ndl, September 2015.

[23] Y. van Pinxteren, G. Geleijnse, and P. Kamsteeg. Deriving a recipe similarity measure for recommending healthful meals. In *Proceedings of the 16th International Conference on Intelligent User Interfaces*, IUI '11, pages 105–114. ACM, 2011.

[24] T. Yamamoto, S. Nakamura, and K. Tanaka. Rerank-by-example: Efficient browsing of web search results. In *International Conference on Database and Expert Systems Applications*, pages 801–810. Springer, 2007.

[25] K. Yanai and Y. Kawano. Food image recognition using deep convolutional network with pre-training and fine-tuning. In *Multimedia & Expo Workshops (ICMEW), 2015 IEEE International Conference on*, pages 1–6. IEEE, 2015.