DOI: 10.15439/2016F579
ACSIS, Vol. 9. ISSN 2300-5963

# Machine Vision in Food Recognition: Attempts to Enhance CBVIR Tools

Andrzej Śluzek

Khalifa University, Abu Dhabi Campus
P.O. Box 127788, Abu Dhabi, UAE
Email: andrzej.sluzek@kustar.ac.ae

***Abstract*—Visual identification of complex images (e.g. images of food) remains a challenging problem. In particular, *content-based visual information retrieval* (CBVIR) methods, which seem a natural choice for such tasks, are often constrained by specific characteristics of the images of interest and (possibly) other practical requirements. In this paper, a novel CBVIR approach to automatic food identification is proposed, taking into account characteristics of solutions currently existing in this area. Based on limitations of those solutions, we present a scheme in which a co-occurrence of MSER features extracted from three color channels is employed to build a *bag-of-words* histogram. Subsequently, food images are matched by detecting similarities between those histograms. Preliminary tests on a recently published benchmark dataset UNICT-FD889 reveal certain advantages of the scheme and highlight its limitations. In particular, a need of a novel methodology for segmentation of food images has been identified.**

## I. Introduction

Unstoppable (and sometimes excessive) presence of mobile devices in everyday activities is understandably followed by development of IT tools which can meaningfully analyze and interpret data collected during those activities. The analysis of visual data is particularly important because such data can be easily an unobtrusively captured (almost) everywhere and (almost) continuously in large quantities. Therefore, there is a growing interest in development of applications for the analysis of diversified categories of everyday-life images and videos. Not surprising, one of attractive and prospectively popular categories is food.

Although it cannot be claimed that automatic recognition of food images/photos becomes a research area of very high importance, growing numbers of publications on this topic can be noticed, e.g. [1], [2], [3], [4], [5], [6], which confirms a certain level of interest.

Some of the presented results (e.g. [2], [7], [8]) target specific health-related applications, i.e. diet assessment, preventing obesity, monitoring food allergies, etc., while others (e.g. [3], [4], [5], [6]) just evaluate applicability of diversified machine vision techniques and algorithms to this particular area.

There are also some works proposing benchmark datasets of food images to evaluate various approaches on those datasets and to stimulate research in the area (e.g. [1], [9].

With the universal availability of smartphones (and more advanced wearable cameras expected in the near future) other practical applications of automatic visual food recognition and identification can prospectively emerge, including quality assessment (serving quality or conformity with presentation standards), search for restaurants serving previously seen dishes, etc.

In this paper, we first (in Section II) briefly overview the most typical approaches to visual food recognition, appraise their advantages, and highlight limitations. In particular, we focus on techniques exploiting the most typical mechanisms of *content-based visual information retrieval* (CBVIR), i.e. local feature (keypoint) detection, description and matching, combined (sometimes)with other algorithms. However, the reported results cannot be considered fully satisfactory. Thus, we attempt to investigate an improved low-level mechanism to enhance reliability of such systems. In Section III, a CBVIR-based approach employing neighborhood dependencies between keypoints extracted from three channels of color images is proposed and preliminarily verified. Unfortunately, the conclusions obtained from the experimental results are not too encouraging either. It seems that food recognition based only on currently existing machine vision techniques cannot reach the level of performances already achieved in other areas of visual data analysis. Apparently, identification of food items using only vision is not as straightforward as expected; some suggestions regarding the future works are discussed in the final Section IV.

## II. Vision-based Techniques for Food Recognition

The earliest attempts to identify food items from their pictorial representations were rather restricted to relatively narrow categories of food represented by individual items, and sometimes discussed from the robotic perspective. For example, a survey of techniques for detecting individual fruits on complicated backgrounds (for machines automatically harvesting fruits) was discussed in [10]. Another example of such a system, i.e. a vision systems for the identification of broken biscuits on a production line was presented in [11].

Those early techniques are typically based on a preliminary detection of predefined shapes (mostly circles or ellipses) followed by the analysis of color and/or texture properties within those extracted shapes. In many cases, the usage of a supplementary range sensors was additionally assumed (driven by the robotic needs). Usually, the realistic scenarios of natural conditions were taken into account, including shadows, unusually bright areas, and object overlapping.

In the following years, more attention was paid to algorithms based on local features and their descriptors (instead of just shape and texture/color characteristics) sometimes supplemented by other, more or less sophisticated image processing tools. One of the best known works based on local features was presented in [9]. The authors combined three algorithms. First, SIFT keypoints, [12], were detected and converted into visual words. Then the whole image was represented by a *bag of visual words* (BoW) histogram, [13], and by another histogram of color distribution. Nevertheless, reliability of this approach was rather unsatisfactory. Even though only a few major ingredients of various fast-food items were considered, accuracy of recognition was below 20% on the dataset of fast-food images defined in the same paper.

The works on the identification of fast-foods were continued in [5] and [6]. In both cases, the same dataset of fast-food images were used. However, the authors of [5] rejected the concept of using SIFT local features or color histograms. Instead, they focused on features characterizing local texture properties of images. The features were build over pairs of pixels (with the feature significance inversely proportional to the distance between pixels). Then, a histogram of pairwise features was the major tool for food detection of eight basic fast-food components (e.g. cheese, bread, egg or beef). Based on detection of those components (and their relative localizations) images were classified into 61 categories, for which the reported accuracy reached 28.2% (for the most successful algorithm of relative localization which supported the basic component identification). When only identification of the seven broader categories (e.g. sandwiches or salads) was considered, the accuracy varied between 69% and 78%.

In [6], the concept of using keypoints and the corresponding visual words was revived. However, instead of SIFT keypoints *local binary patterns* (LBP, [14]), which were considered more suitable for texture characterization than SIFTs, were used. The accuracy for the seven broad categories of fast-foods varied from 56% to 90%. The authors also found that for some categories performances of the method based on SIFT keypoints were superior to LBP.

Altogether, even for a narrow category of fast-food items only there is no clear picture regarding the recommended techniques for visual identification of food. The situation is more complicated if a wider range of food items is to be considered. For example, in [4], 85 diversified items of Japanese food were analyzed. The author used a fusion of several image features, i.e. SIFT-based BoWs, Gabor features and color histograms, and applied multiple-kernel learning techniques to achieve accuracy exceeding 60%.

In [3], a wider dataset of 100 Japanese dishes was considered, in which each dish may contain two or more food items (e.g. fish and chips with salad). The food images were preliminarily segmented into individual items using trained classifiers to detect and evaluate segmentation regions, and each detected region was recognized by a multiple-kernel learning technique. Again, usefulness of SIFT features and color histograms was acknowledged there, but they were combined with HoG and (similarly to [4], Gabor texture features.

Currently, it seems the most comprehensive (in terms of the number of food items) study was presented in [1]. UNICT-FD889 dataset of almost 900 diversified dishes (see examples in Fig. 1) have been collected and used to compare and benchmark performances of food recognition algorithms based on various representation models.

Fig. 1. Exemplary images of UNICT-FD889 dataset (from [1])

The major conclusion was that CBVIR approaches (i.e. methods based on keypoint-like feature detection and image matching) are, in general, applicable to food images. Three types of feature-based representations were eventually selected, namely SIFT, Bag of Textons ( [15]) and pairwise rotation invariant co-occurrence linear binary patterns ( [16]).

Bag of Textons was preliminarily found superior, but the other two representations were not far behind (especially if applied in the variants intended for color images). Nevertheless, the authors did not apply any trained classifier. Therefore, the results were generally inferior to those reported in the previously discussed papers. However, such an approach seems more practical since it would be impossible to have a classifier for each newly encountered food category (e.g. dishes seen the first time).

## III. Matching Food Images

### A. Methodologigal principles

In this paper we propose a novel method for matching food images, that is conceptually more similar to [1] rather than to the other papers discussed in Section II. Therefore, no classifiers have been built for the known food categories, and the method is open to unknown types of food without any modification or retraining . In general, we assume that:

1) Images are represented by collections of local features which are subsequently converted (through quantization of their descriptors into *visual words*) to *bags of words* (BoW) histograms.

2) The features represent images in a wider visual context, i.e. feature descriptors characterize features in conjunction with a number of neighboring features. Moreover, the features are separately extracted from individual color channels so that those semi-local characteristics incorporate the color properties as well.
3) At the current level of development, images are compared using only BoW histograms, but the future developments may incorporate more sophisticated usage of the features and their descriptors.

Actually, the selected features are MSER keypoints (see [17]) which have low complexity and good performances. They are usually represented by elliptic approximations so that features of elongated shapes (which are frequently present in images of food) can be more accurately handled (compared, for example, to shapes represented by SIFT keypoints).

MSER features are separately found in three color channels (i.e. R, G and B images) and, subsequently, they are analyzed semi-locally using the method similar to the approach preliminarily outlined in [18] and applied in a more sophisticated version in a number of later works (e.g. [19], [20]).

Thus, for any MSER feature of *C* color (where *C* = R, G or B) represented by $E_C$ ellipse we define its *S*-color neighborhood as a collection of *S*-color MSERs (where *S* = R, G or B, and $S \neq C$) as follows:

An *S*-color MSER represented by $E_S$ ellipse belongs to *S*-color neighborhood of $E_C$ ellipse if:

1) The distance $d(E_C, E_S)$ between origins of $E_C$ between $K$ and $E_S$ satisfies:

$$1/2 \times r_{norm} \leq d(E_C, E_S) \leq 2 \times r_{norm}, \qquad (1)$$

where $r_{norm} = \sqrt{area(E_C)/\pi}$.

2) The areas of $E_C$ and $E_S$ ellipses are similar but $E_C$ is larger (i.e. the ratio is between 0.5 and 1).

Using a large collection of images (including a significant percentage of food images) we have found that the average size of such neighborhoods is between 8 and 10.

In each color channel, individual MSER features (i.e. their ellipses) are represented by SIFT descriptors in RootSIFT variant which has been found superior (see [21]). Subsequently, RootSIFT descriptors of the keypoint ellipses $E$ are quantized into visual words $w(E)$ from a vocabulary of either 128 or 1024 (two variants have been implemented) words.

Then, for each keypoint with $E_C$ ellipse we take into account its two *S*-color neighborhoods (containing a number of $E_S$ ellipses). For example, if $C = R$ the neighborhoods will be built using MSER keypoints from *green* and *blue* channels.

Eventually, pairs of visual words $w(E_C)$ and $w(E_S)$ are formed, and each such a pair is described by a word from a vocabulary of either $128 \times 128 = 64k$ or $1024 \times 1024 = 1M$ words. Thus (because the average size of neighborhoods is between 8 and 10) each keypoint of *C*-color contributes, in average, $16 - 20$ visual words to the *bag-of-words* (BoW) histogram of the image.

Finally, similarities between images are estimated by the similarities between their BoW histograms built according to the above principles.

Because our approach is not restricted to images of known and predictable food items, BoW normalization techniques requiring database statistics (e.g. *td-idf*, [13]) cannot be applied, and we use histograms of *absolute* word frequencies in images.

Numerous measures of histogram similarities exist (e.g. [22]) but not all of them are applicable to BoW matching. Because of the assumptions applied in this work during BoW building, we eventually selected a simple *histogram intersection* measure (proposed in [23]), where the distance between two histograms $H_A$ and $H_B$ over *Voc* vocabulary is defined by

$$d(H_A, H_B) = \sum_{w \in Voc} min(H_A(w), H_B(w)). \qquad (2)$$

Such a measure nicely corresponds to the intuitive notion of similarity between both full images and sub-images (including textured images).

### B. Preliminary experimental results

The proposed approach was verified on UNICT-FD889 dataset discussed earlier. Not all dishes were fully tested, but we focused primarily on two most typical cases of food images. First, plates filled by uniformly looking dishes were considered (see examples in Fig. 2). Secondly, dishes represented by images with a number visually different regions of various food components (see Fig. 3) were taken into account.

The retrieval performances for both categories of dishes have been found very different. For uniform dishes, the top retrievals are usually highly relevant images. Images most similar to query images from Fig. 2 are shown in Figs 4 and 5, correspondingly. They highly correspond to the human perception, even though in Fig. 4 two categories of foods are mixed up.

For dishes consisting of several non-uniformly distributed items, performances are rather miserable. A spectacularly incorrect example is given in Fig. 6.

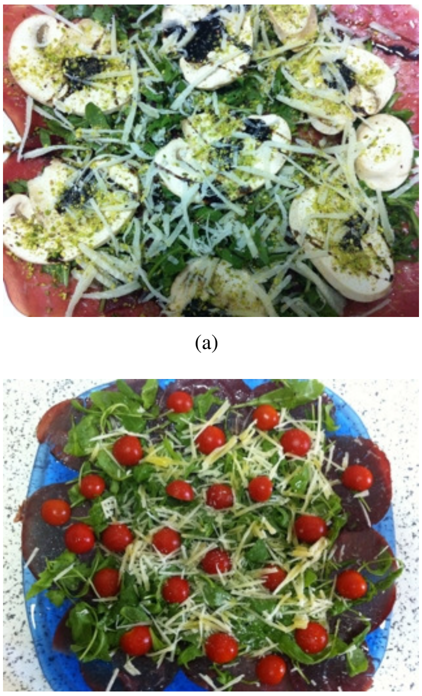
However, there are also cases (an example given in Fig. 7) when search results are quite good for multiple-item dishes.

## IV. Concluding remarks

The paper proposes a novel CBVIR-based scheme for visual identification of food. Using the (apparently) largest publicly available dataset, we have tested a method based on BoW histograms which actually represent semi-local co-occurrences of features (MSER keypoints are selected as examples) extracted from three color channels of RGB images. Similar image retrieval is based on the similarities between such histograms.

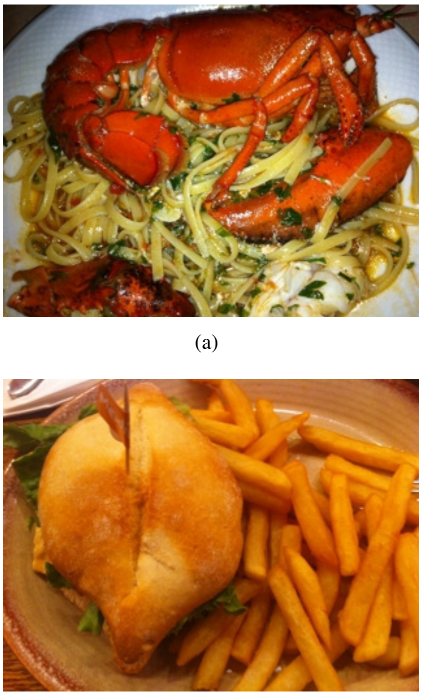
Because of it low complexity, the scheme could be considered an attractive option for limited-performance mobile devices equipped with a camera.

Unfortunately, the experimental verification has been found only partially successful. The results are satisfactorily accurate


(a)

(b)

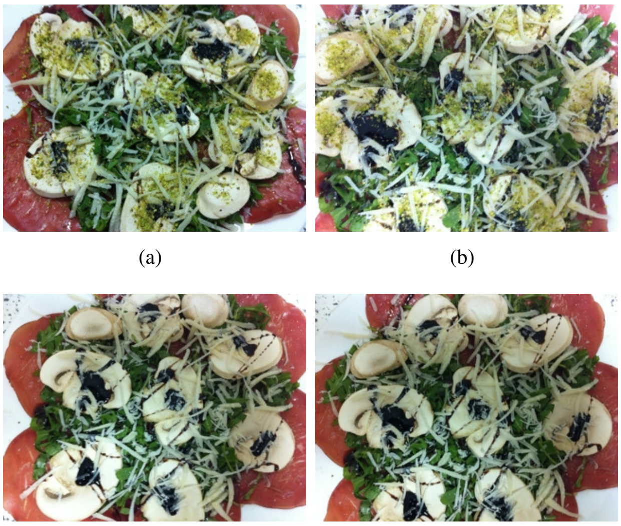
Fig. 2. Examples of dishes uniformly filling plates with similarly looking contents (from UNICT-FD889 dataset).


(a)

(b)

Fig. 3. Examples of dishes consisting of diversified components (from UNICT-FD889 dataset).

only for dishes looking uniformly over the whole plate. For mixtures of diversified foods shown on the same plate,


(a)

(b)

(c)

(d)

Fig. 4. Top retrievals for the image from Fig. 2a. Note that (a,b) and (c,d) are actually considered different dishes.

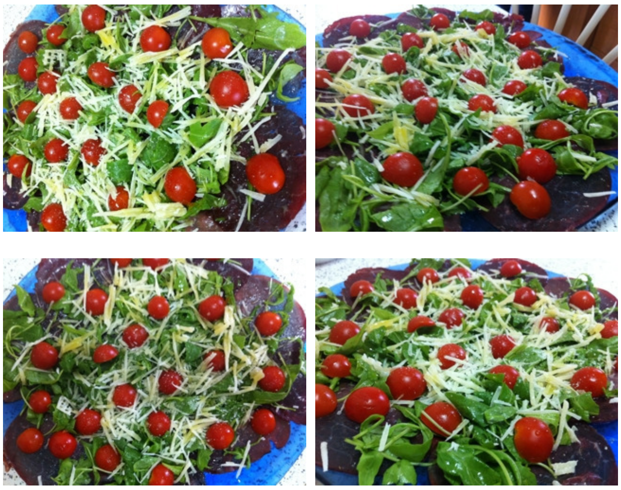


Fig. 5. Top retrievals for the image from Fig. 2b.

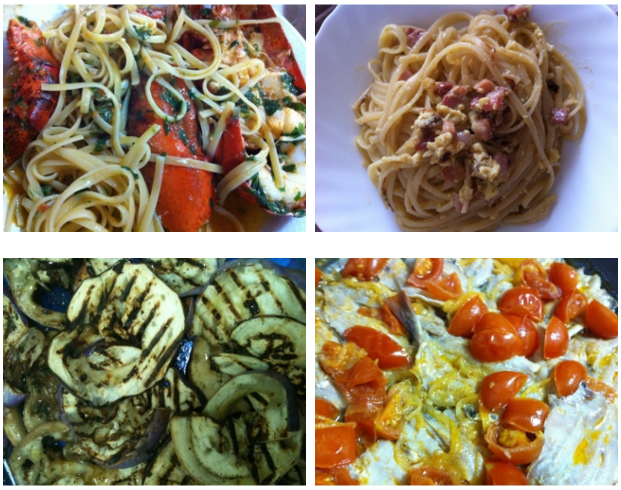


Fig. 6. Top retrievals for the image from Fig. 3a.

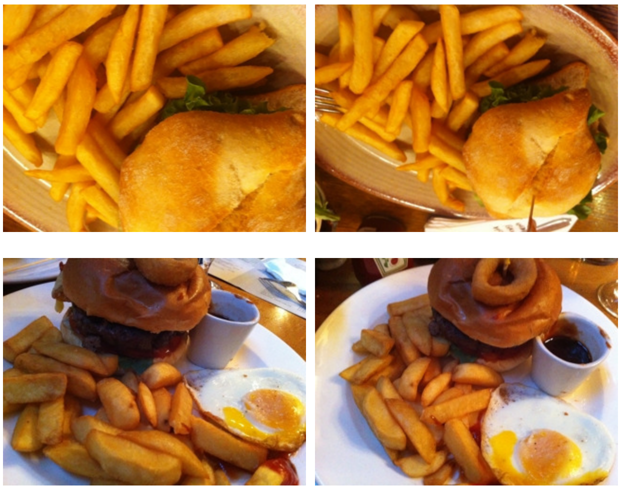


Fig. 7. Top retrievals for the image from Fig. 3b.

performances are unacceptably low (although in some cases perfomances are acceptable). Therefore, it can be preliminarily concluded that the approaches presented in some of the previous works, which require image segmentation into uniform regions (e.g. [3]) have been validated (in terms of the proposed methodologies).

However, those segmentation techniques proposed in the published works only partially correspond to needs identified in our experiments. For example, dishes shown in Fig. 2 should be considered uniform regions, but the existing segmentation technique (even if incorporating texture-based approaches) would apparently not segment them in the required way.

Altogether, wec can conclude that fully automatic vision-based food identification still remains a challenging problem. The main challenge is apparently segmentation of multiple-item dishes into uniform region, which can be prospectively recognized using keypoint-based approaches. Nevertheless the satisfactory solutions for such segmentation have not been identified yet. Keypoint-based co-segmentation (e.g. [24]) is one of the most promising approaches.

## References

[1] G. M. Farinella, D. Allegra, and F. Stanco, "A benchmark dataset to study the representation of food images," in *Proc. ECCV 2014 Workshops*, vol. III, 2015, pp. 584–599. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-16199-0_41

[2] F. Kong and J. Tan, "Dietcam: Automatic dietary assessment with mobile camera phones," *Pervasive and Mobile Computing*, vol. 8, no. 1, pp. 147–163, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.pmcj.2011.07.003

[3] Y. Matsuda, H. Hoashi, and K. Yanai, "Recognition of multiple-food images by detecting candidate regions," in *Proc. IEEE Int.Conf. on Multimedia and Expo*, 2012, pp. 25–30. [Online]. Available: http://dx.doi.org/10.1109/ICME.2012.157

[4] H. Hoashi, T. Joutou, and K. Yanai, "Image recognition of 85 food categories by feature fusion," in *Proc. IEEE Int. Symposium on Multimedia*, 2010, pp. 296–301. [Online]. Available: http://dx.doi.org/10.1109/ISM.2010.51

[5] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, "Food recognition using statistics of pairwise local features," in *Proc. IEEE Conf. CVPR 2010*, 2010, pp. 2249–2256. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2010.5539907

[6] Z. Zong, D. Nguyen, P. Ogunbona, and W. Li, "On the combination of local texture and global structure for food classification," in *Proc. IEEE Int. Symposium on Multimedia*, 2010, pp. 204–211. [Online]. Available: http://dx.doi.org/10.1109/ISM.2010.37

[7] G. O'Loughlin, S. Cullen, A. McGoldrick, S. O'Connor, R. Blain, S. O'Malley, and G. Warrington, "Using a wearable camera to increase the accuracy of dietary analysis," *American Journal of Preventive Medicine*, vol. 44, no. 3, pp. 297–301, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.amepre.2012.11.007

[8] F. Zhu, M. Bosch, I. Woo, S. Kim, C. Boushey, D. Ebert, and E. Delp, "The use of mobile devices in aiding dietary assessment and evaluation," *Journal of Selected Topics in Signal Processing*, vol. 4, no. 4, pp. 756–766, 2010. [Online]. Available: http://dx.doi.org/10.1109/JSTSP.2010.2051471

[9] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, and J. Yang, "Pfid: Pittsburgh fast-food image dataset," in *Proc. IEEE Conf. ICIP 2009*, 2009, pp. 289–292. [Online]. Available: http://dx.doi.org/10.1109/ICIP.2009.5413511

[10] A. Jimenez, A. Jain, R. Ruz, and J. Rovira, "Automatic fruit recognition: a survey and new results using range/attenuation images," *Pattern Recognition*, vol. 32, no. 10, pp. 1719–1739, 1999. [Online]. Available: http://dx.doi.org/10.1016/S0031-3203(98)00170-8

[11] F. Pla, "Recognition of partial circular shapes from segmented contours," *Comput. Vision & Image Understanding*, vol. 63, no. 2, pp. 334–343, 1996. [Online]. Available: http://dx.doi.org/10.1006/cviu.1996.0023

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004. [Online]. Available: http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94

[13] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Conf. ICCV 2003*, vol. 2, Nice, 2003, pp. 1470–1477. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2003.1238663

[14] T. Ojala, M. Pietikainen, and D. Harwood, "A comparative study of texture measures with classification based on feature distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996. [Online]. Available: http://dx.doi.org/10.1016/0031-3203(95)00067-4

[15] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *International Journal of Computer Vision*, vol. 62, no. 1-2, pp. 61–81, 2005. [Online]. Available: http://dx.doi.org/10.1007/s11263-005-4635-4

[16] X. Qi, R. Xiao, J. Guo, and L. Zhang, "Pairwise rotation invariant co-occurrence local binary pattern," in *Proc. ECCV 2012*, 2012, pp. 158–171. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-33783-3_12

[17] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, pp. 761–767, 2004. [Online]. Available: http://dx.doi.org/10.1016/j.imavis.2004.02.006

[18] C. Schmid and R. Mohr, "Local grayvalue invariants for image retrieval," *IEEE Trans PAMI*, vol. 19, no. 5, pp. 530–535, 1997. [Online]. Available: http://dx.doi.org/10.1109/34.589215

[19] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," in *Proc. IEEE Conf. CVPR 2009*, 2009, pp. 25–32. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2009.5206566

[20] A. Śluzek, "Extended keypoint description and the corresponding improvements in image retrieval," *LNCS (Revised Selected Papers of ACCV 2014 Workshops)*, vol. 9008, pp. 698–707, 2015. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-16628-5_50

[21] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Proc. IEEE Conf. CVPR 2012*, 2012, pp. 2911–2918. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2012.6248018

[22] S.-H. Cha and S. Srihari, "On measuring the distance between histograms," *Pattern Recognition*, vol. 35, pp. 1355–1370, 2002. [Online]. Available: http://dx.doi.org/10.1016/S0031-3203(01)00118-2

[23] M. Swain and D. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991. [Online]. Available: http://dx.doi.org/10.1007/BF00130487

[24] A. Śluzek and M. Paradowski, "Reinforcement of keypoint matching by co-segmentation in object retrieval: Face recognition case study," *LNCS (Proc. ICONIP 2012)*, vol. 7667, pp. 34–41, 2012. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-34500-5_5