Autocropping: A Closer Look at Benchmark Datasets

Luigi Celona[ORCID], Gianluigi Ciocca[✉][ORCID], Paolo Napoletano[ORCID], and Raimondo Schettini[ORCID]

Department of Informatics, Systems and Communication,
University of Milano-Bicocca, Viale Sarca, 336, 20126 Milan, Italy
{luigi.celona,gianluigi.ciocca,paolo.napoletano,
raimondo.schettini}@unimib.it

Abstract. Image cropping is a common image editing task that aims to improve the composition as well as the aesthetics of an image by extracting well-composed sub-regions of the original image. For choosing the "best" autocropping method it is therefore important to consider on which datasets this method is validated and possibly trained. In this work we conduct a detailed analysis of the main datasets in the state of the art in terms of statistics, diversity and coverage of the selected sub-regions, namely the ground-truth candidate views. An analysis of how much semantics of ground-truth candidate views is preserved with respect to original images and a comparison among dummy autocropping solutions and state of the art methods is also presented and discussed. Results show that each dataset models the cropping problem differently, and in some cases very high performance can be reached by using a dummy autocropping strategy.

Keywords: Autotagging · Dataset comparison · Evaluation

1 Introduction

Image cropping is an important step to improve the aesthetic quality of an image. It can be used in many applications such as efficient image transmission, image retargeting, or photo collages [1]. For choosing the "best" image cropping method it is important to consider on which datasets this method is validated and possibly trained. Computer Vision (CV) researchers benchmark algorithms using public available datasets [11]. This practice allows to achieve quantitative performance evaluation and comparison between algorithms as well as it helps new researchers to get, in a short time, a clear view of the state of the art performance. The availability of benchmark datasets is increasing in all the CV domains ranging from object tracking [15] to object recognition [7]. Although much effort has been made to enrich the number of available datasets, in contrast, a little effort has been made to assess the quality of the available ones. Quality is related to two aspects: (i) how much data are representative of the domain; (ii) how much the ground-truth is consistent with the modeled problem.

In this paper we provide a thorough analysis of datasets that are commonly used for the benchmark of autocropping algorithms. In the last 10 years several benchmark datasets have been proposed by the scientific community as well as suitable measures to rank algorithms in terms of match between ground-truth sub-region or sub-regions with the one or ones selected by the algorithm [2, 6, 21, 22]. Each dataset usually consists in a set of images and corresponding cropped sub-region or sub-regions as selected by human subjects. In order to evaluate the characteristic and effectiveness of the datasets in the literature, in this paper, we analyze the most used datasets in terms of: (1) statistics of the ground-truth, namely position, size and aspect-ratio of the candidate views; (2) diversity and coverage of each candidate views with respect to the original image; (3) performance evaluation and comparison with the state of the art of a dummy solution that crops regions with area ranging from 100% to 10% of the image area; (4) semantic analysis in terms of number of times a semantic concept is preserved in ground-truth sub-regions with respect to original images.

2   Related Works

2.1   Datasets for Image Cropping Assessment

Many databases for the evaluation of image autocropping methods have been proposed in the literature. Table 1 summarizes the available datasets. For each dataset we report the number of images it contains, the number of views per image i.e. the number of crops available for each image, the source of the images, how the crops have been determined (either by human annotation, or by an automatic procedure), who validated the crops i.e. if the human subjects were experts in the field or not, whether the different views are ranked by preference, and finally the corresponding reference where the dataset has been presented for the first time. Briefly, the characteristics of the five datasets in Table 1 are the following.

Table 1. Comparison of publicly available image cropping databases.

Dataset	Images	Views	Source	Crops	Evaluation	Ranking	Ref
CPC	10,797	24	Misc[a]	Generated	AMT workers	Yes	[21]
CUHK-ICD	950	3	CUHKPQ	Human	Experts	No	[22]
FCDB	348	1	Flickr	Human	AMT workers	No	[2]
FLMS	500	10	Flickr	Human	Experts	No	[6]
XPView	992	≤ 23	Misc[b]	Generated	Experts	Yes	[21]

[a]AVA, MS-COCO, AADB, Places
[b]FCDB, CUHK-ICD, MS-COCO,...

Comparative Photo Composition Database. The Comparative Photo Composition (CPC) database contains 10,797 images [21]. For each image, 24 candidate views with 4 standard aspect-ratios have been pooled among candidates automatically generated by exploiting existing re-composition and cropping algorithms. Finally, the aforementioned candidate views have been ranked by 6 Amazon Mechanical Turk (AMT) workers. The source of the images is quite diverse. It consists of a combination of images taken from different benchmark datasets in the literature: AVA [16], MS-COCO [13], AADB [10] and the Places dataset [24]. Most of the images contain two or more principal objects.

CUHK Image Cropping Database. The CUHK Image Cropping Database (CUHK-ICD) is a collection of 950 images gathered from the CUHKPQ dataset [22]. It contains seven classes of images, i.e. *animal*, *architecture*, *human*, *landscape*, *night*, *plant* and *static*. A cropped region is respectively annotated for each image by three different professional photographers. The images are taken from an existing image quality assessment dataset, the CUHKPQ dataset [18]. The images are of varying aesthetic quality and are of different image categories.

Flickr Cropping Database. The Flickr Cropping database (FCDB) contains 1,743 non-iconic images gathered from Flickr [2]. The cropping annotation for each image derives from the choices of four AMT workers who evaluated several candidate views manually drawn. 348 out of the 1,743 images are adopted as test set and is the dataset's cardinality reported in Table 1. Also, since there are no multiple views for each image, and thus no ranking of different crops, in the table the "Ranking" attribute is set to "No" for this dataset.

FLMS Database. The FLMS database consists of 500 images crawled from Flickr [6]. These images have been selected for their imperfect composition and have different contents. Each image is cropped by 10 expert users on AMT who passed a strict qualification test. There is no ranking of the views. Each view is considered separately. No further details are provided in [6] about this dataset.

eXPert View Database. The eXPert View (XPView) database is a collection of 992 images with dense compositions [21]. This dataset has been created by the same authors of the CPC dataset in order to test their method on another, unrelated dataset. The origin of the XPView is mixed with the images taken from different sources. Specifically, the MS-COCO [13], the FCDB and CUHK-ICD datasets, and other, unspecified, sources. The candidate views have been generated as already described for the CPC dataset but with 8 diverse aspect-ratios. In this case, the candidate views are annotated by three experts, and a ranking of the views is provided. From the analysis of the dataset, each image has up to 23 views.

## 2.2   Image Cropping Algorithms

The problem of automatic image cropping has been traditionally tackled by designing ad-hoc algorithms based on different visual cues that are considered

relevant. Many methods consider salient regions to guide the selection of the important portion of the images. For example, in [4], depending on the image contents, different cropping attributes are used such as faces, skin, saliency and the image category itself. Another example is [14]. Visual composition, boundary simplicity and content preservation are used in [6] as features to force a cropped image to contain a salient object. Yan et al. [22] propose a method for learning what features are important in a good crop among color, texture, foreground, shape complexity, sharpness, saliency maps, segmented regions, perspective ration, and prominent lines.

Recently, another category of cropping algorithms has emerged that incorporate aesthetic cues as a feature in order to select the best cropping region. For example, a Generative Adversarial Network is used in [5] with a discriminator that attempts to distinguish images of poor and good aesthetic quality. Aesthetic and gradient energy maps are used in [9] to learn a compositional model for the best crop. The View Finding Network (VFN) [3] tries to correctly rank candidate crops according to certain photographic guidelines learned on an aesthetically annotated database. In [19,20] candidate crops are firstly generated and then their aesthetic is assessed to generate the cropped image. A similar approach based on aesthetic quality classification is the CNN-based Cascaded Cropping Regression (CCR) method [8]. Li et al. [12] propose an Aesthetic Aware Reinforcement Learning (A2-RL) framework to sequentially search the best cropping windows automatically generated by applying a set of cropping actions. Finally, a fast View Proposal Net (VPN) is presented in [21], where a teacher network, is used to teach the VPN (i.e. the student) to output the correct score rankings for the crops.

3 Autocropping Ground-Truth: Candidate Views Analysis

Figure 1 shows some sample images generated from each dataset with the corresponding candidate views superimposed. For the CPC and the XPView datasets, the views are ranked so greenish colors represent the best ranked views while the reddish colors represent the worst ranked ones. As it can be seen, the views selected for both the CPC and XPView have a high degree of variability with the different views covering the most part of the original image. For the CUHK-ICD, we can see that the three views mostly overlap although in some cases they can be quite different (as in the case of the building and the cake). The single, favorite, candidate view of the FCDB dataset covers the relevant object in the image. Given the presence in the image of a single relevant subject, the images themselves seem quite simple to crop. Finally, for the FLMS dataset, again, the ten candidate views are quite diverse. For instance, the bounding boxes have small overlaps. This could indicate that the ten experts have different personal opinions about image aesthetics and composition. Following the above preliminary examination, we next analyze in details the five datasets, their annotations, and provide some observations on their use for the evaluation of automatic cropping algorithms.

CPC

CUHK-ICD

FCDB

FLMS
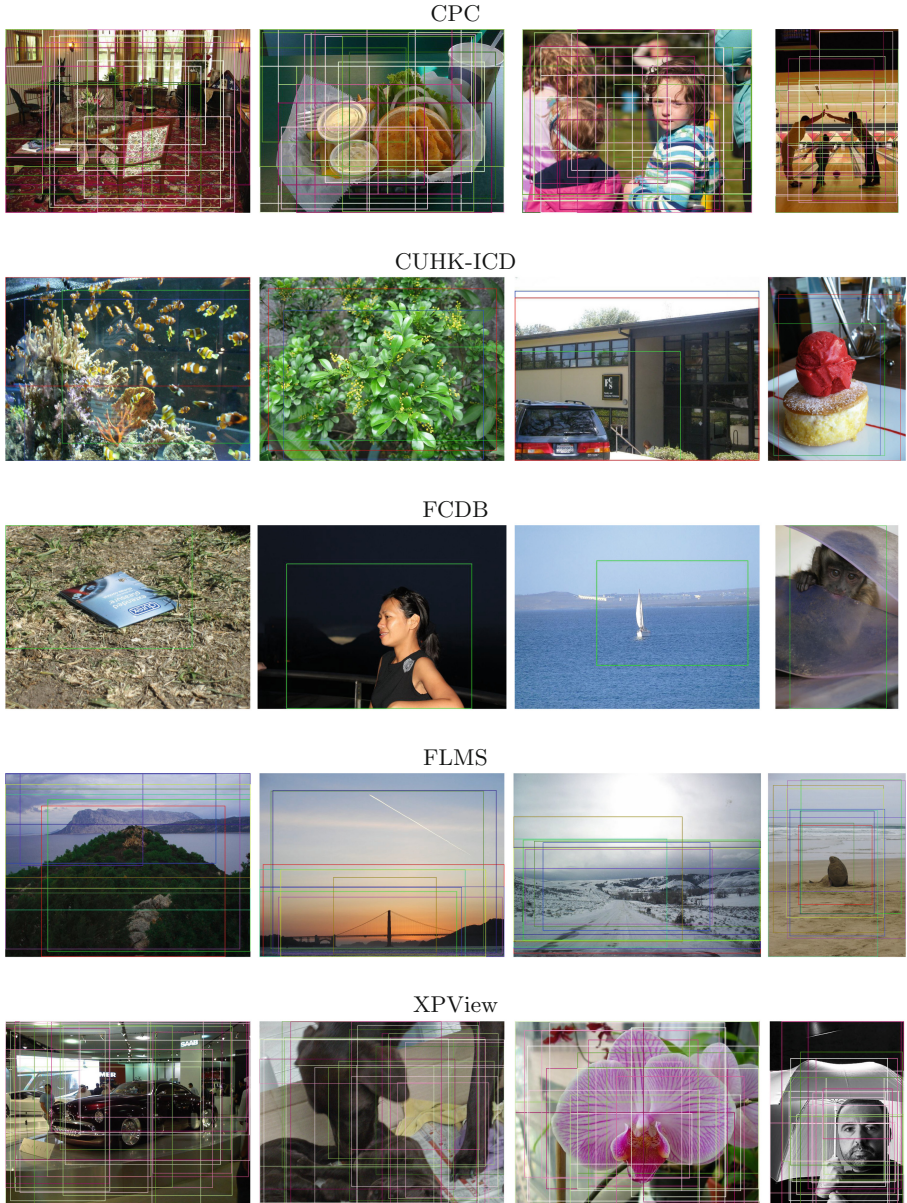
XPView

Fig. 1. Sample images from the five cropping datasets. Superimposed are the crop regions. For the CPC and the XPView datasets, reddish to greenish colors represent ranked crops from worst to best. (Color figure online)
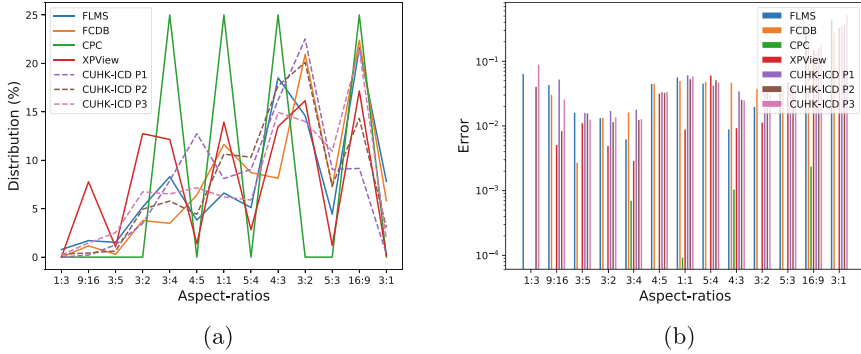
Fig. 2. *(Best viewed magnified.)* Distribution of candidate views for each considered dataset with respect to 13 aspect-ratios commonly used in digital photography (a). Average error between aspect-ratios of the candidate views and closest standard aspect-ratios for each database (b).

3.1   Diversity and Coverage

For each dataset, we quantitatively analyze several properties of candidate views. Firstly, we investigate the aspect-ratio of the candidate views to understand how much these differ from common aspect-ratios. We categorize candidate views aspect-ratios into 13 common classes in still camera photography, namely 1:1, 5:4, 4:3, 3:2, 5:3, 16:9, 3:1 and their complementary versions (4:5, 3:4, 2:3, 3:5, 9:16, and 1:3) [23]. Figure 2a shows the distribution of candidate views aspect-ratios for each database. As it can be seen, the majority of candidate views for all the datasets has a 16:9 aspect-ratio. CPC dataset candidate views equally distribute among 1:1, 4:3, 16:9, and 3:4 aspect-ratios. The other datasets (CUHK-ICD, FCDB, FLMS, and XPView) have a larger variety of candidate views aspect-ratios. Figure 2b reports the error resulting from the categorization step, that is the average distance between candidate views aspect-ratios and the closest standard aspect-ratios for each dataset. The small error for the CPC dataset is motivated by the fact that candidate views were also sampled from standard aspect-ratios as described in Sect. 2.1. Instead, the error for the other datasets is higher because candidate views have been freely chosen by humans.

Secondly, we consider the surface of all candidate windows for estimating their diversity and also their coverage with respect to the surface of the whole image. We scale the size of images as well as the corresponding candidate views to the same fixed dimension, then the value of the pixel o_{ij} of the heatmap, which represents the probability of being part of a candidate view, is obtained as follows:

$$o_{ij} = \frac{1}{N} \sum_{n=1}^{N} \alpha(x_{ij}) \qquad \alpha = \begin{cases} val, & \text{if } x_{ij} \in W_n \\ 0, & \text{otherwise} \end{cases} \qquad (1)$$

where N is the total number of candidate views for all dataset samples, W_n is the candidate n-th view and val corresponds to 1 for datasets that do not provide the rank of candidate views, namely CUHK-ICD, FCDB, and FLMS dataset. For CPC and XPView datasets, whose candidate views do not have all the same relevance, val is equal to w_n, where it represents the rank of the n-th candidate view, normalized in the interval [0,1]. In Fig. 3 we display the heatmaps for all the datasets. The heatmaps are normalized in the range [0,1], where pixel value close to 1 means that there is a high probability that the corresponding pixel belongs to a candidate view. The high energy in the center of the heatmaps for all the datasets shows that many candidate views crop the central region of the image. Moreover, we highlight that the energy is very high for almost the entire surface of CUHK-ICD images, while it is lower in the edges of the other datasets, in particular, those of the CPC and XPView. This means that most of the CUHK-ICD candidate views cover almost the total surface of the image, while the CPC and XPView candidate views are very different from each other and focus on image regions much smaller than the entire surface. The previous qualitative results are validated by quantitative analysis. Precisely, we estimate the average percentage of candidate views coverage respect to the whole image area. The values obtained for each photographer P of the CUHK-ICD dataset correspond respectively to: 82.07 ± 14.74 for P1, 82.69 ± 17.89 for P2, and 80.49 ± 16.76 for P3. For FCDB dataset, it is equal to 65.55 ± 16.64. The percentage coverage obtained for the candidate views of the FLMS dataset is 58.59 ± 17.41. Finally, CPC and XPView datasets have similar statistics equal respectively to 41.82 ± 16.95 and 43.68 ± 17.07.
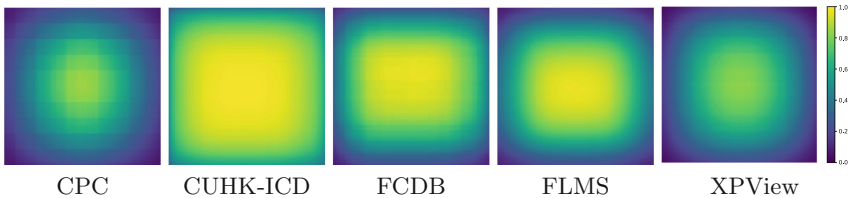

Fig. 3. Heatmaps showing the spatial coverage of all candidate views for each dataset.

Finally, we analyze the semantic of images and how the crop of candidate view sub-regions alter it. To this end, we exploit the Hybrid-CNN [24], a CNN trained using 3.5 million images for 1,183 categories, obtained by merging the scenes categories from Places database [24] and the object categories from ImageNet [17]. The table in Fig. 4a reports the percentage of times that the semantic concept of an image is maintained in the crop obtained by applying the candidate views. Figure 4b presents the distribution of semantic concepts on the images of each dataset. We can see that the distributions are very spread across all the categories and some peaks are present in correspondence of landscape concepts like: promontory, lakeside, and valley.

Dataset		Same concept (%)
CPC		38.89
	P1	65.89
CUHK-ICD	P2	68.21
	P3	68.00
FCDB		50.87
FLMS		58.94
XPView		39.89

(a)
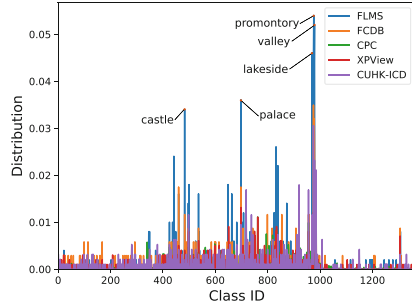
(b)

Fig. 4. Semantic analysis. Number of times semantic concept is preserved in candidate views for each dataset (a). Distribution on the 1,365 semantic concepts of the Hybrid-CNN for dataset images (b).

3.2 Performance Evaluation of Autocropping Algorithms

We measure the baseline by considering a dummy solution consisting of crops sampled in different ways with a surface that covers the image area with decreasing percentages from time to time. More in detail, we estimate the performance by cropping regions keeping from 100% to 10% of the image area, and by averaging the results of 100 iterations of random crops retaining from 100% to 10% of the image area. The evaluation metrics commonly used for cropping performance comparison are intersection-over-union (IoU) and boundary displacement error (BDE).

Intersection-over-Union (IoU). The intersection-over-union (IoU), also referred to as the Jaccard index, is essentially a method to quantify the percent overlap between the ground-truth candidate view and the predicted crop. Given the area of the ground-truth candidate view W_{GT} and the area of the predicted crop W, the IoU is defined as follows:

$$IoU = (W_{\mathrm{GT}} \cap W)/(W_{\mathrm{GT}} \cup W) \tag{2}$$

Boundary Displacement Error (BDE). The boundary displacement error computes the distance between the four edges of the ground-truth candidate view and the corresponding edges of the predicted crop. By denoting the four edges of the ground-truth candidate view and of the predicted view respectively as $B_{\mathrm{GT}}(l)$, $B_{\mathrm{GT}}(r)$, $B_{\mathrm{GT}}(t)$, $B_{\mathrm{GT}}(b)$, and $B(l)$, $B(r)$, $B(t)$, $B(b)$. The BDE is estimated as follows:

$$BDE = \sum_{j=\{l,r,u,b\}} |B_{\mathrm{GT}}(j) - B(j)|/4, \tag{3}$$
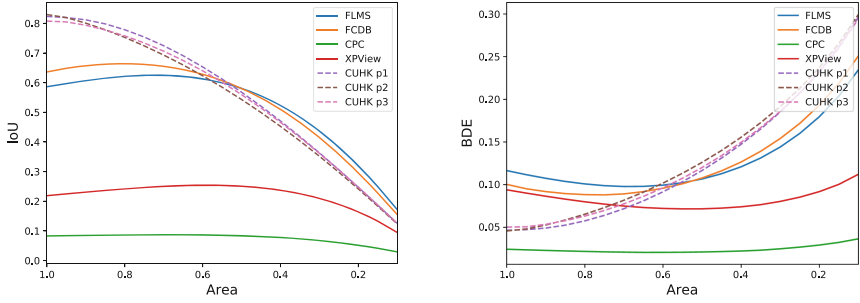
Fig. 5. IoU and BDE obtained by comparing ground-truth candidate views with dummy central crops covering image areas from 100% to 10%.

Table 2. Comparison in terms of IoU and BDE respectively for CUHK-ICD, FCDB and FLMS datasets. The $\Delta Best$ row reports the absolute error between the result (in blue) of the best method in the state of the art and the best value (in magenta) among the various scales of the dummy solution.

Method	CUHK-ICD						FCDB		FLMS	
	P1		P2		P3					
	IoU	BDE	IoU	BDE	IoU	BDE	IoU	BDE	IoU	BDE
VFN (2017) [3]	0.749	0.071	0.729	0.075	0.732	0.074	0.675	0.086	0.747	0.067
A2-RL (2018) [12]	0.802	0.052	0.796	0.053	0.790	0.053	0.663	0.089	0.820	–
ABP-AA (2017) [19]	0.813	0.030	0.806	0.032	0.816	0.032	–		0.810	0.057
AIC (2018) [20]	0.815	0.031	0.810	0.030	0.830	0.029	0.650	0.080	0.830	0.052
CCR (2018) [8]	0.850	0.032	0.837	0.033	0.828	0.035	–		–	
VPN (2018) [21]	–		–		–		0.711	0.073	0.835	0.044
VEN (2018) [21]	–		–		–		0.735	0.072	0.836	0.041
Dummy (100%)	0.823	0.046	0.830	0.046	0.808	0.050	0.636	0.100	0.586	0.116
Dummy (95%)	0.819	0.047	0.819	0.048	0.805	0.050	0.648	0.095	0.597	0.112
Dummy (85%)	0.798	0.052	0.778	0.058	0.778	0.058	0.661	0.089	0.615	0.104
$\Delta Best$	0.027	0.015	0.007	0.016	0.022	0.021	0.074	0.017	0.221	0.063

Results. We collect performance for all the datasets at varying crop scales both for the center and random dummy solutions: the two dummy solutions achieved performance that is not significantly different. Figure 5 exhibits the IoU and the BDE at varying center crop scales for each database. As it is possible to see, the performance for the CUHK-ICD dataset is initially very high, both in terms of IoU and BDE, and declines in an almost linear fashion as the surface covered by the dummy crops decreases. Achieved results for FLMS and FCDB are linear until scale 0.5 where they go down. Finally, performance is stably low for CPC and XPView datasets. Table 2 shows comparison, in terms of IoU and BDE, between several algorithms in the state of the art with the dummy solution. We include CUHK-ICD, FCDB and FLMS because they are the datasets commonly used for benchmarking cropping algorithms. From the table is clear that CCR [8] and AIC [20] are the best methods in terms of IoU and BDE for the CUHK-ICD dataset. However, the best dummy solution achieves, on the same dataset,

an IoU that is about 2.7% lower and a BDE that is about 1.5% lower than the best in the state of the art. This behavior can be explained by looking at the heatmap of the CUHK-ICD dataset (see Fig. 3). The heatmap shows that the ground-truth candidate views cover quite completely all the image. VEN [21] algorithm is the best on FCDB and FLMS datasets. In the case of FCDB, the best dummy solution achieves a performance that is 7.4% and 1.7%, in terms of IoU and BDE, lower than the best in the state of the art. In the case of FLMS, the best dummy solution achieves a performance that is 22.1% and 6.3% lower than the best in the state of the art. FLMS dataset contains ground-truth candidate views at aspect-ratios that are quite different from the common ones (see Fig. 2b). Moreover, Fig. 3 shows that candidate views do not cover the entire image.

4   Conclusions

In this work we conduct a detailed analysis of the main datasets in the state of the art for the evaluation of autocropping methods in terms of statistics, diversity and coverage of the ground-truth crops. Results show that each dataset models the cropping problem differently. Moreover, CPC and XPView datasets consist of very diverse candidate views, and most of the datasets do not consist of candidate views having standard aspect-ratios. Comparison between state of the art and dummy solutions show that, in case of the CUHK-ICD dataset, comparable results with the best solution in state of the art can be reached by using a dummy autocropping strategy that does not crop anything. Results obtained on the FCDB and FLMS show that these datasets are more challenging and diverse, with the dummy solutions performing worse than state of the art algorithms, and thus making them more suitable for the evaluation of autocropping algorithms.

References

1. Bianco, S., Ciocca, G.: User preferences modeling and learning for pleasing photo collage generation. ACM TOMM **12**(1), 6 (2015)
2. Chen, Y.L., Huang, T.W., Chang, K.H., Tsai, Y.C., Chen, H.T., Chen, B.Y.: Quantitative analysis of automatic image cropping algorithms: a dataset and comparative study. In: WACV, pp. 226–234. IEEE (2017)
3. Chen, Y.L., Klopp, J., Sun, M., Chien, S.Y., Ma, K.L.: Learning to compose with professional photographs on the web. In: ICM, pp. 37–45. ACM (2017)
4. Ciocca, G., Cusano, C., Gasparini, F., Schettini, R.: Self-adaptive image cropping for small displays. IEEE TCE **53**(4), 1622–1627 (2007)
5. Deng, Y., Loy, C.C., Tang, X.: Aesthetic-driven image enhancement by adversarial learning. In: ICM, pp. 870–878. ACM (2018)
6. Fang, C., Lin, Z., Mech, R., Shen, X.: Automatic image cropping using visual composition, boundary simplicity and content preservation models. In: ICM, pp. 1105–1108. ACM (2014)
7. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: The KITTI vision benchmark suite (2015). http://www.cvlibs.net/datasets/kitti

8. Guo, G., Wang, H., Shen, C., Yan, Y., Liao, H.Y.M.: Automatic image cropping for visual aesthetic enhancement using deep neural networks and cascaded regression. IEEE Trans. Multimed. **20**(8), 2073–2085 (2018)
9. Kao, Y., He, R., Huang, K.: Automatic image cropping with aesthetic map and gradient energy map. In: ICASSP, pp. 1982–1986. IEEE (2017)
10. Kong, S., Shen, X., Lin, Z., Mech, R., Fowlkes, C.: Photo aesthetics ranking network with attributes and content adaptation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 662–679. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_40
11. Kotsiantis, S., Kanellopoulos, D., Pintelas, P., et al.: Handling imbalanced datasets: a review. GESTS Int. Trans. Comput. Sci. Eng. **30**(1), 25–36 (2006)
12. Li, D., Wu, H., Zhang, J., Huang, K.: A2-RL: aesthetics aware reinforcement learning for image cropping. In: CVPR, pp. 8193–8201. IEEE (2018)
13. Lin, T.Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
14. Marchesotti, L., Cifarelli, C., Csurka, G.: A framework for visual saliency detection with applications to image thumbnailing. In: ICCV, pp. 2232–2239. IEEE (2009)
15. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: MOT16: a benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016)
16. Murray, N., Marchesotti, L., Perronnin, F.: AVA: a large-scale database for aesthetic visual analysis. In: CVPR, pp. 2408–2415. IEEE (2012)
17. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015)
18. Tang, X., Luo, W., Wang, X.: Content-based photo quality assessment. IEEE Trans. Multimed. **15**(8), 1930–1943 (2013)
19. Wang, W., Shen, J.: Deep cropping via attention box prediction and aesthetics assessment. In: CVPR, pp. 2186–2194. IEEE (2017)
20. Wang, W., Shen, J., Ling, H.: A deep network solution for attention and aesthetics aware photo cropping. IEEE TPAMI **41**, 1531–1544 (2018)
21. Wei, Z., et al.: Good view hunting: Learning photo composition from dense view pairs. In: CVPR, pp. 5437–5446. IEEE (2018)
22. Yan, J., Lin, S., Kang, S.B., Tang, X.: Learning the change for automatic image cropping. In: CVPR, pp. 971–978. IEEE (2013)
23. Zhang, M., Zhang, L., Sun, Y., Feng, L., Ma, W.: Auto cropping for digital photographs. In: ICME, pp. 4-pp. IEEE (2005)
24. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Advances in Neural Information Processing Systems, pp. 487–495 (2014)