Fine-grained Evaluation on Face Detection in the Wild

Bin Yang* Junjie Yan* Zhen Lei Stan Z. Li
Center for Biometrics and Security Research & National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, China
{yb.derek,yanjjie}@gmail.com {zlei,szli}@nlpr.ia.ac.cn

***Abstract*— Current evaluation datasets for face detection, which is of great value in real-world applications, are still somewhat out-of-date. We propose a new face detection dataset MALF (short for Multi-Attribute Labelled Faces), which contains 5,250 images collected from the Internet and ∼12,000 labelled faces. The MALF dataset highlights in two main features: 1) It is the largest dataset for evaluation of face detection in the wild, and the annotation of multiple facial attributes makes it possible for fine-grained performance analysis. 2) To reveal the 'true' performances of algorithms in practice, MALF adopts an evaluation metric that puts stress on the recall rate at a relatively low false alarm rate. Besides providing a large dataset for face detection evaluation, this paper also collects more than 20 state-of-the-art algorithms, both from academia and industry, and conducts a fine-grained comparative evaluation of these algorithms, which can be considered as a summary of past advances made in face detection. The dataset and up-to-date results of the evaluation can be found at http://www.cbsr.ia.ac.cn/faceevaluation/.**

I. INTRODUCTION

Face detection plays an important role in face based image analysis and is one of the fundamental problems in computer vision. The performances of various face based applications, from traditional face identification and verification to modern face clustering, tagging and retrieval, rely on accurate and efficient face detection. Popular detectors, such as Viola-Jones detector [25] and its subsequences (e.g., vector boosting [11]) have achieved satisfactory performance on early datasets, such as CMU-MIT. However, as argued in modern works [12], [29], the Viola-Jones based methods are still far from prefect. Many new detection-related methods have been proposed recently, such as DPM [6], CNN [8] and multiple channel features [1], which have been proven or suggested to be able to improve performance of face detection, such as [27], [26], [20], [28]. Besides these academic researches, face detection is also put great efforts to by commercial companies, such as Google, Facebook and Face++. Among all the above approaches, however, we do not know which one is the best and how to improve them for real world scenarios. Therefore, a well-designed benchmark is in urgent demand to clear up the confusion and push forward the progress of face detection. Unfortunately, we still lack an unbiased real world face detection benchmark for the following three reasons.

The first is that current face detection benchmarks do not support the fine-grained analysis of detection results, which makes the quantitative exploration of causes and the correlation between different types of errors difficult. Current face detection benchmarks only have the rough bounding box (or estimated fitting ellipse like FDDB [12]) annotations, and can only report an overall face detection result on the whole test set. As we know, the poses, glasses, expressions can influence the detection result considerably, but the current face detection benchmarks cannot tell us how much these influences are and which one is the more important factor than another. Commonly, one algorithm may have its advantage on some conditions and disadvantage on others, thus roughly reporting an overall performance would take the risk of ignoring the strong point of the algorithm.

The second is that current face detection benchmarks do not reflect the 'true' real world. The most widely used face detection benchmark in early years is the MIT+CMU test set. The testing images are all of gray scale and collected in ten years ago, and there exists a large domain gap with current web images due to the technical progress in digital cameras. Another widely used face detection benchmark is the FDDB [12], which is collected from news photographs. However, these faces tend to be salient in the image, and the pose tends to be frontal. For more widely used consumer images, such as the images from Flickr, Facebook and Google+, the faces can be more diverse. Recently, a large scale face database named AFLW [14] is released with detailed landmark annotations. However, some of the faces in the database are not annotated, making it less suitable to serve as a face detection benchmark than as a training set.

The third is that current face detection benchmarks do not report the 'true' state-of-the-art results. Current face detection works often only compare the performance with academic algorithms on MIT+CMU and FDDB. However, as pointed in [31], there is a large gap between currently available academic solutions and commercial systems (e.g. Google Picasa) and online commercial API (e.g. Face++). Comparisons with pure academic methods cannot guarantee high quality in real world applications.

This paper addresses the above-mentioned problems by making the following contributions:

1) We collect a large face dataset for face detection evaluation. The database scale is currently the largest among face detection test sets. We also annotate multiple attributes along with bounding box of faces in the dataset. As far as we know, it is the first time that such a large face dataset is thoroughly labelled (Table I).

2) We propose a fine-grained evaluation methodology based on multi-attribute annotations, by defining test-set-of-interest with attribute labels. We give a straightforward

* indicates co-first authorship.

example in the evaluation part of this paper by defining 'easy' sub-set, 'moderate' sub-set and 'hard' sub-set. The flexible evaluation methodology is able to reflect the advantages and disadvantages of the evaluated algorithms with regard to diverse facial attributes, like pose, gender, resolution, wearing glasses and so on.

3) We evaluate 21 state-of-the-art face detection algorithms both from industry and academia. Some of them are submitted by authors upon our request, some are implemented using open source codes, and some results of commercial systems are counted by person. We conduct fine-grained evaluation on all these algorithms and provide analysis on the results. Our thorough evaluation and analysis could give insights into where we should focus our efforts for further improvements.

The remaining of this paper is organized as follows: the next sub-section briefly introduces the related work in evaluation dataset and methodologies. Section 2 describes our dataset, including the data collection and annotation guidelines as well as statistical properties. Section 3 explains the fine-grained evaluation methodology. Evaluation and performance analysis of the state-of-the-arts are shown in Section 4. In the last section we conclude the paper.

A. Related Work

In this part we review some related and remarkable evaluation datasets, covering tasks of both object detection and face detection.

The most influential challenge in this decade may be the Pascal VOC Challenge [5]. The challenge consists of classification, detection and segmentation. In detection task, the dataset contains annotations of objects in 20 different classes, while each annotation includes not only the bounding box coordinates, but also the following attributes: 'orientation', 'occluded', 'truncated' and 'difficult'. These attributes are specified for selective training and 'ignore' flag during evaluation. In the competition, the challenge organizers introduce a number of novel evaluation methods, like Boostrapping AP and rank, and normalized precision for cross-class comparison [4]. Recently, the ImageNet challenge [22] largely extends Pascal VOC by incluing more categories and more images. The Caltech Pedestrian Dataset [2] is widely used for pedestrian detection. Considering the special case of pedestrian detection in the vehicle view, Piotr carefully designs the guidelines in data collection and annotation strategy. For example, he adopts per-image evaluation rather than per-window evaluation, and labels both the visible and whole extent of the person while the whole extent is used for evaluation.

As for the face detection, the most frequently used evaluation datasets are AFW [31] and FDDB [12]. The AFW dataset contains 205 images collected from Flickr with 468 labelled faces. Annotations include a rectangular bounding box, 6 landmarks and the pose angles. The FDDB dataset contains 2845 images with 5171 faces, while each face is annotated with a pre-defined ellipse instead of bounding box. Both the images database and annotations of these two datasets are released and researchers can conduct the evaluation themselves. Performance are ranked according to the plotted curves (Precision-Recall curve in AFW and ROC curve in FDDB) on the whole test set.

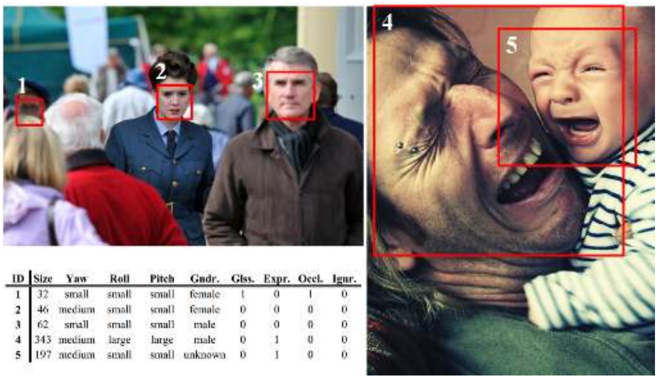
TABLE I
Comparison Of Test Sets For Face Detection In The Wild

Dataset	#Img	#Face	Property	Annotation
CMU/MIT testset[24], [21]	125	483	gray-scale frontal	6 landmarks
CMU profile [23]	208	441	gray-scale frontal&profile	6/9 landmarks
AFW [31]	205	468	color in the wild	rect. box 6 landmarks view angle
FDDB [12]	2,846	5,171	gray&color in the wild	bounding ellipse
MALF	5,250	11,931	color in the wild	square box 5 attributes

[4] states that in the current multi-category object detection evaluation setting, the diversity of state-of-the-art algorithms is limited because a novel method may not beat a quite mature conventional method in evaluation performance. This problem also exists in face detection evaluation. However, we argue that by adopting the fine-grained evaluation, which could evaluate the attribute-specific performances of algorithms, novel method may stand out in one or two sub-set evaluations, which therefore testifies its novelty in some certain aspects. [10] identifies different types of errors occurred in object detection, which gives useful advice on how to improve the performance. R. Benenson [20] points out that specifically in face detection domain, the evaluation is usually unfair due to different bounding box policies used by different datasets and scale difference between dataset annotation and detector output.

II. DATASET

MALF dataset contains in total 5,250 high-resolution images from the Internet. The images are collected in the following two steps: 1) About 2,000 images are manually collected from Flickr, and around 30,000 images are collected using the similar image search service provided by Baidu Inc. to guarantee that most of them contain people. 2) All images are then manually examined by two persons to pick out images that are included in the dataset. The selection procedure follows the principle to guarantee large diversity in face appearances. There are eventually 5,250 images included in the MALF dataset, containing in total 11,931 labelled faces in the wild. The dataset size is the largest among currently available face detection evaluation datasets. Among all 5,250 images, we randomly take out 250 images as example images. Algorithm designers can use the annotations of example images to do transfer learning or adjust the output bounding box style of their algorithm. The rest 5,000 images are purely test images. Like AFW and FDDB datasets, we don't provide an individual training set, as some algorithms require large amount of data, and some even require landmark annotations.



ID	Size	Yaw	Roll	Pitch	Gndr.	Glss.	Expr.	Occl.	Igur.
1	32	small	small	small	female	1	0	1	0
2	46	medium	small	small	female	0	0	0	0
3	62	small	small	small	male	0	0	0	0
4	343	medium	large	large	male	0	1	0	0
5	197	medium	small	small	unknown	0	1	0	0

Fig. 1. Example images and annotations in the dataset.

A. Multi-attribute Annotation

For all 5,250 images in the MALF dataset, the bounding box of all recognizable faces as well as a boolean ‘ignore’ flag are first labelled. The bounding box is an axis-aligned square in similar style to that in AFLW dataset. Specifically, the bounding box tries to contain the eyebrow, the chin and the cheek, while keeping the nose located approximately at the box center (see Fig. 1 for an example). ‘Ignore’ flag is set to true if the face is very difficult to recognize due to very large occlusion, blurring and other extreme deformations, or the size of bounding box is below 20 (totally 838 faces, account for round 7%). In order to keep annotations consistent, the bounding box is annotated by two persons and examined by one, and the ‘ignore’ flag is annotated by one person.

After the initial annotation step, for each face with false ‘ignore’ flag, we further annotate the following attributes: gender (male, female, unknown), pose deformation level of yaw, pitch and roll (small, medium, large), occluded (true/false), wearingGlasses (true/false), exaggeratedExpression (true/flase). We don’t label specific pose angles due to the large workload in landmark annotation and pose estimation. Instead, we define three levels of pose deformation and convert the pose annotation into a classification problem. Each attribute is annotated by one of the two persons and then examined by the other in order to keep attribute definition consistent in the dataset.

B. Dataset Statistics

In this part we present the statistical properties of MALF dataset and its annotations. See Fig. 1 for an example. In terms of the image data, all collected images are RGB images and in JPEG format. The average image size is 573 pixels high and 638 pixels wide. Each image contains 2.27 faces in average, with 46.97% of images contain one face, 43.41% contain 2∼4 faces, 8.30% contain 5∼9 faces, and 1.31% images contain more than 10 faces.

In terms of face annotations, distribution of each attribute of face is shown in Fig. 2. As for the face scale, the mean size of all faces is 83×83, and the median face size is 64×64 large. We choose the size of 60 and 90 to divide the scale range into small, medium and large intervals. Note than faces smaller than 20×20 are always labelled as ‘ignore’ (account for 5.88% of all faces).

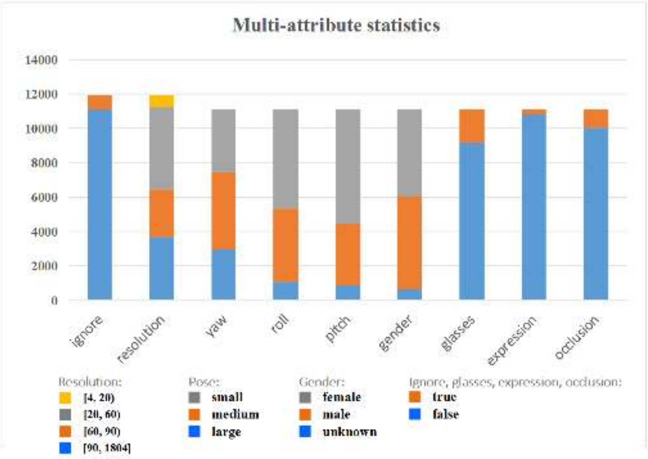


Fig. 2. Statistics of multiple attributes in the dataset.

III. EVALUATION METHODOLOGY

Our evaluation differs from other face detection evaluations in two main points. The first is the fine-grained evaluation protocol, which is feasible thanks to the multi-attribute annotations. The second is a stress on performance at low false alarm rate by plotting curves in the log space of FPPI (False Positive Per Image) rate.

We first describe our detection evaluation rules for clarification (Algorithm 1). Similar to PASCAL VOC Challenge, we first rank all detection results in descending order according to their corresponding confidence scores. For each detection, we find the ground-truth which shares the largest IoU (Intersection over Union) ratio with the detection itself. If it is found, and the IoU ratio is larger than a threshold (which is 0.5 is our case), then we decide whether it is a true positive or a false positive according to whether the found ground-truth has been detected before. Note that ‘ignore’ face is an exception. Miss/correct/multiple detection of it won’t be counted. This Image-based evaluation requires a post processing of detection results in an image to remove multiple detections of the same face. As we can see from the algorithm, multiple detection of one face will increase the FPPI but won’t change the True Positive Rate. We leave it to the algorithm designers themselves and do not modify their algorithms’ outputs. Performance is represented via ROC curve by varying the confident score threshold from high to low, with TPR (True Positive Rate) being the y axis and FPPI (False Positive Per Image) in log scale being the x axis.

A. Fine-grained Evaluation

The additional annotation of facial attributes makes the fine-grained evaluation of face detection possible. Theoretically, we can report specific performance with regard to gender, pose, resolution (size of faces), glasses, occlusion and expression respectively. However, in practice, as some

Algorithm 1 Evaluation Rules

```
Rank $Scores$, $Detections$ in score-descending order.
Set the $ignore$ flag of $gt$ in $GTs$ whose attribute labels don't fit the sub-set definition as true.
for $i$ = 1:length($Scores$) do
    $Dets$ = $Detections$[1:$i$];
    $tp$ = 0; $fp$ = 0;
    Set $detected$ flag in whole $GTs$ as false;
    for each $BB\_det \in Dets$ do
        [$BB\_gt$,$IoU$]=findMaxOverlapGt($BB\_det$,$GTs$);
        if $IoU > threshold$ then
            if $BB\_gt.ignore$ == false then
                if $BB\_gt.detected$ == true then
                    $fp$++;
                else
                    $tp$++;
                    $BB\_gt.detected$ = true;
                end if
            end if
        else
            $fp$++;
        end if
    end for
    $tpr[i]$ = $tp$ / #(GTs with false ignore flag);
    $fppi[i]$ = $fp$ / #Images;
end for
```

attribute labels are quite sparse and some are highly correlated with one another (e.g. faces wearing sun glasses are occluded), evaluations on attribute-specific sub-set may be biased. Following the principles in another object detection dataset, KITTI [7] benchmark, which defines three levels of difficulty, we define 'easy' and 'hard' sub-sets with different combinations of attribute labels. Concretely, 'easy' sub-set contains faces larger than 60×60, without any large pose, occluded or exaggerated expression; 'hard' sub-set contains faces larger than 60×60, with one of the extreme conditions (large pose, occluded or with exaggerated expression). Besides these two pre-defined sub-sets, algorithm designers can also conduct attribute-specific evaluations to further analyze the performance using the multiple attribute annotations.

In evaluation of each sub-set, ground truths which don't belong to this sub-set are marked as 'ignore' temporarily, therefore the evaluation can reflect the performance with regard to only the considered attribute. Note that the fine-grained evaluation is not constrained to one attribute only, but also supports any combination of labelled attributes. This means that, the fine-grained evaluation is highly customized and could provide more thorough performance analysis than before.

B. Evaluation Metric

In face detection evaluation, two curves are frequently used, i.e., Precision-Recall curve and ROC curve, and two numeric metrics are also widely employed, which are Average Precision value and Area Under Curve value. As a rare-case problem, in real-world applications, what we care about is how the algorithm performs (the recall rate) at a high precision level (low false alarm). However, current face detection evaluation curves and metrics lose the most valuable information very much. Most of these metrics highlight high recall rate with less attention to the precision. Following the method used in [2], we plot the True Positive Rate - False Positive Per Image curve in the log space of FPPI. FPPI is an appropriate measurement of precision in per-image object detection and the log scale stresses the performances at low FPPI rates. For example, in video surveillance applications where number of targets is large in each image, TPR at 10^0 FPPI may be suitable for algorithm evaluation; while in handful device based applications, TPR at 10^{-2} FPPI would be more appropriate. For performance comparison between different algorithms, we also define a numeric metric like the mean-miss rate defined in [2], which we call it mean-recall rate. The mean-recall rate is calculated as the average true positive rate at 9 evenly sampled points between 10^{-2} FPPI and 10^{-1} FPPI in log space[1]. The higher this value is, the better the performance is.

IV. EVALUATION OF THE STATE-OF-THE-ART

To make MALF a meaningful benchmark, we collect 21 face detection algorithms from academia and industry to be evaluated. The specific methods we collect and comparative performance results of them are presented in this section. Analysis on the evaluation results are discussed as well.

A. Methods

21 face detectors, including 5 commercial systems, are evaluated on the MALF dataset. As for commercial softwares (Google Picasa, Apple iPhoto and Windows Photo Gallery), we don't have their specific bounding box results. Instead, we manually count the true positives and false positives on the whole test set. Therefore they only support evaluation on whole test set, while don't support the proposed fine-grained evaluation[2]. As for Face++, we refer to its online free API as v1, and refer to the Face++ internal version's results submitted by its authors as v2. As for the rest 16 academic algorithms, they are either submitted by their authors, or implemented using open source codes and/or models provided by the original authors. We also collect training data information and implementation parameters for an all-round description. All these detailed information of the 21 algorithms, including authors, institution, codename for the entry, are listed in Table II. For all evaluated algorithms, there are several main approaches deserving notice. Specifically, deformable part model is quite popular in object detection, while ID 7 and 12 belong to this category. Viola-Jones framework is a classic method and ID 9, 13, 14, 16, 17, 19 and 21 belong to this type. Channel features

[1] Note that if the algorithm outputs only bounding box results without scores, the performance curve becomes a single point in the figure and the mean-recall rate cannot be calculated.

[2] Google Picasa is an exception as we manually count its performance on each fine-grained evaluation conducted in this paper for the sake of a strong baseline in each evaluation.

ID	CodeName	Author(s)	Institution(s)	Training Data	Source	Parameters
1	iPhoto	–	Apple	–	iPhoto	Version 9.6
2	FacePP_v1	–	Megvii	–	Online Free API	Request date: 2014.10.08
3	FacePP_v2	–	Megvii	–	Submission	–
4	Picasa	–	Google	–	Google Picasa	Version 3.7
5	Gallery	–	Microsoft	–	Windows Photo Gallery	Version 16.4.3528.331
6	ACF	B. Yang, J. Yan, Z. Lei, S.Z. Li [28]	CBSR & NLPR, Chinese Academy of Sciences	AFLW database	Submission	multi_scale feature, 6 views, imresize 3x
7	DPM	M. Mathias, R. Benenson M. Pedersoli, L. Van Gool [20]	iMinds & MPI Informatics	AFLW, Pascal Face dataset	Open source model	threshold = -0.5 imresize 2x
8	Exemplar	H. Li, Z. Lin, J. Brandt, X. Shen, G. Hua [16]	Stevens Institute of Tech. & Adobe Research	15, 832 face images 12, 732 non-face images	Submission	–
9	VJ-hyb	Y. Gavini	VIT University	–	Submission	Viola-Jones based hybrid detector
10	Headhunter	M. Mathias, R. Benenson M. Pedersoli, L. Van Gool [20]	iMinds & MPI Informatics	AFLW, Pascal Face dataset	Open source software Doppia	Headhunter model
11	ICF	M. Mathias, R. Benenson M. Pedersoli, L. Van Gool [20]	iMinds & MPI Informatics	AFLW, Pascal Face dataset	Open source software Doppia	Headhunter_baseline model
12	B·DAT	J. Deng, J. Yang, D. Wang, S. Yan, G. Liu, Q. Liu	NUIST	AFLW, additional data set ∼1 million faces	Submission	Improved DPM, context, alignment, imresize 2x
13	Pico	N. Markus, M. Frljak, I. S. Pandzic, J. Ahlberg, R. Forchheimer [19]	University of Zagreb	∼20k frontal face images	Submission Codes available on GitHub	scale factor: 1.075, stride factor: 0.05
14	NPD	S. Liao, A.K. Jain, S.Z. Li [18]	CBSR & NLPR, Chinese Academy of Sciences	FDDB	Submission	–
15	SPM	Ahmed EL-Barkouky, Ahmed Shalaby, Ali Mahmoud, Aly Farag [3]	CVIP Lab, University of Louisville	Helen [15] & FDDB	Submission	–
16	SurfCas	J. Li, T. Wang, Y. Zhang [17]	Intel Labs China	–	Open source codes	model_type = 1, minsz = 8
17	SZU	S. Yu	Shenzhen University	–	Submission	–
18	TSM	X. Zhu, D. Ramanan [31]	University of California, Irvine	Multi-PIE [9]	Open source codes	face_p146_small, threshold = -2
19	VJ	P. Viola, M.J. Jones [25]	Microsoft Research, Redmond Mitsubishi Electric Research Laboratory	–	Open source software OpenCV	haarCascade models: frontalface_default & profileface
20	W.S.Boost	Z. Kalal, J. Matasm, K. Mikolajczyk [13]	University of Surrey Czech Technical University	–	Open source codes	models: frontal & profile
21	MBLBP	L. Zhang, R. Chu, S. Xiang, S.Z. Li [30]	CBSR & NLPR, Chinese Academy of Sciences	–	Submission	–

TABLE II
DETAILED INFORMATION OF EVALUATED FACE DETECTION ALGORITHMS

is a new feature representation used to improve the Viola-Jones framework, while ID 6, 10 and 11 use this approach. ID 3 adopts the recently well-known Convolutional Neural Network approach. Fig. 5 shows some detection results of these three categories of algorithms.

B. Results and Analysis

We first show the evaluation results on the whole test set (see Fig. 3). From the curve, we can see that among commercial systems, FacePP_v2 and Picasa achieve outstanding performances with very high TPR at low FPPI. iPhoto shows competitive results at very low FPPI (actually the number of false positives of iPhoto is zero). FacePP_v1 and Gallery get relatively poor results. As for the academic algorithms, it is hard to say which one performs better than another from the curve. In terms of recall rate, the B-DAT owns a quite strong edge over others. However, it gets poor performance at lower FPPI. In terms of overall performance, the ACF and SZU get top results.

The smallest face in MALF is 20×20, which is quite challenging for many algorithms. Although many evaluated algorithms have upscaled the test images, the scale factor still affects the performance very much. Therefore, we conduct a scale-related evaluations, with ‘small’ corresponds to size smaller than 60×60 and ‘large’ corresponds to size larger than 90×90. Seen from the curves in Fig. 3, in ‘small’ subset, while performances of all algorithms drop a little, the performance of DPM based methods (DPM and B-DAT) degrades more compared to other approaches, like channel features based methods. One possible explanation would be that the feature representation used in DPM (usually HoG) works relatively poor in low resolution scenarios, while the channel features own the property of scale approximation [1]. Nevertheless, in ‘large face’ subset, the situation changes over. DPM based methods have a considerable performance boost over channel features based methods. Considering the difference between these two types of approaches, it could be inferred that this time without the feature depressing caused by low resolution, DPM which explicitly models the structure constraints achieves a higher recall rate than channel features based models which implicitly models the structure information.

Here we conduct fine-grained evaluations on two defined sub-sets, ‘easy’ and ‘hard’ (see Section III.A for definitions).

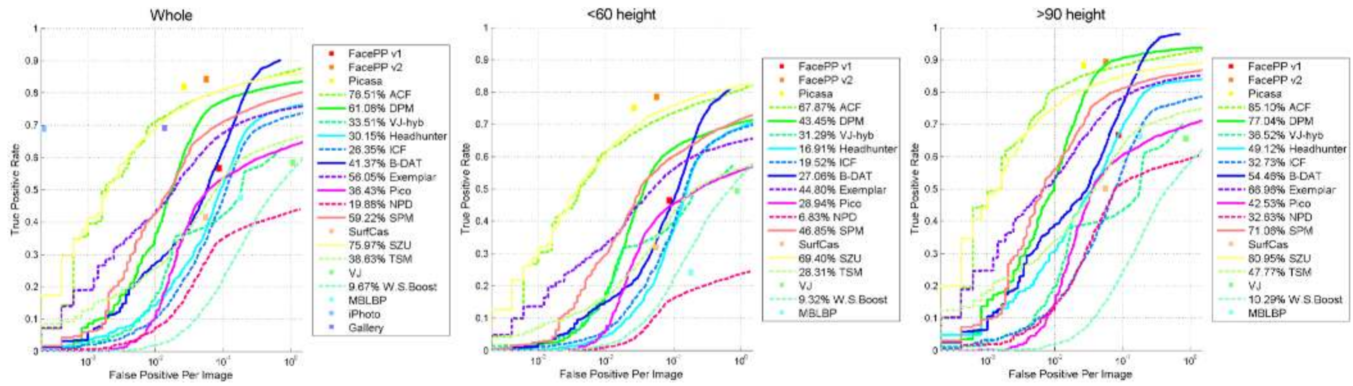


Fig. 3. Fine-grained evaluation on the whole test set, small faces sub-set and large faces sub-set.

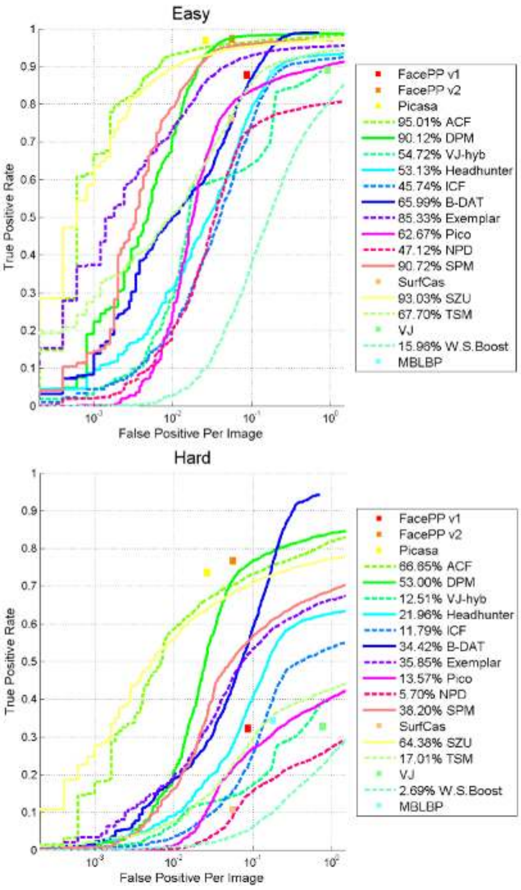


Fig. 4. Fine-grained evaluation on 'easy' and 'hard' faces sub-sets.

Note than both sub-sets only take faces larger than 60×60 to remedy the side effects caused by small faces. From the curves illustrated in Fig. 4, we can see that on the 'easy' sub-set, when the precision is low, i.e., FPPI larger than 10^{-1}, almost all algorithms achieve a recall rate over 80%, some even over 95%, therefore it's pointless in comparing performances at a low level of precision on easy faces. Instead, it should be more appropriate to measure the TPR when the FPPI is smaller than 10^{-2}. Here ACF and SZU perform well under such settings. When it comes to 'hard' sub-set, the performance diversity becomes much larger. By observing the performances at high level of precision, for example, when FPPI is lower than 10^{-2}, the best academic algorithm achieves less than 60% TPR and it decreases dramatically as precision level rises.

V. DISCUSSION AND CONCLUSION

In this paper, we propose a new face detection dataset MALF (short for Multi-Attribute Labelled Faces) with annotations of multiple attributes. MALF features a fine-grained evaluation methodology with a stress on algorithm performance at high precision level. However, there are still some problems in the current version of this evaluation dataset. First is the imperfect annotations. As argued in [20], different bounding box styles in different datasets makes it unfair to directly matching algorithm outputs with the labelled ground-truth, especially when the ground-truth contains background. The attribute labels also have room for improvements as currently most attributes are boolean. Second is the detection evaluation criterion. The IoU overlap threshold 0.5 may be too arbitrary and too loose for real-world applications, a threshold of 0.7 may be more appropriate. These are the directions in which we are moving forward.

In summary, with the fine-grained evaluation, we can analyze the performance of the algorithm in different aspects with regard to multiple attributes. By comparing performances in varying precision levels, we can further observe the advantages and disadvantages of the algorithm in various scenarios. With these two components combined, MALF could serve as a helpful face detection benchmark which offers deep and all-round diagnosis and improvement advice on evaluated algorithms.

VI. ACKNOWLEDGMENTS

We'd like to give our sincere gratitude to the valuable advice given by all the reviewers, as well as Rodrigo Benenson and Shiqi Yu. This work was supported by the Chinese National Natural Science Foundation Projects #61203267, #61375037, #61473291, National Science and Technology Support Program Project #2013BAK02B01, Chinese Academy of Sciences Project No. KGZD-EW-102-2 and AuthenMetric R&D Funds.

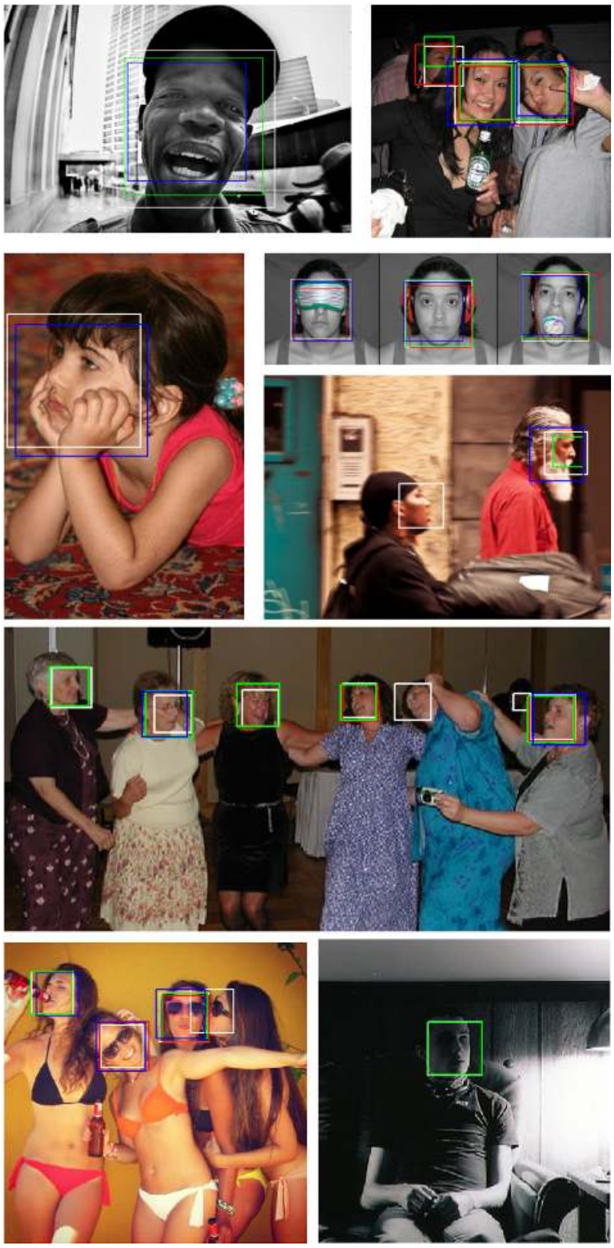


Fig. 5. Ground-truth (white) and detection results of 'FacePP_v2' (red), 'ACF' (green) and 'DPM' (blue) on some test images. Faces without bounding box of specific color are missed by the corresponding algorithm (displayed at a score threshold of 10^{-1} FPPI on the whole test set). Best viewed in color on screen.

References

[1] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. 2014.

[2] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *PAMI*, 34, 2012.

[3] A. EL-Barkouky, A. Shalaby, A. Mahmoud, and A. Farag. Selective part models for detecting partially occluded faces in the wild. In *ICIP*. IEEE, 2014.

[4] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge–a retrospective.

[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.

[7] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv preprint arXiv:1311.2524*, 2013.

[9] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *IVC*, 2010.

[10] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *Computer Vision–ECCV 2012*, pages 340–353. Springer, 2012.

[11] C. Huang, H. Ai, Y. Li, and S. Lao. High-performance rotation invariant multiview face detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(4):671–686, 2007.

[12] V. Jain and E. G. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. *UMass Amherst Technical Report*, 2010.

[13] Z. Kalal, J. Matas, and K. Mikolajczyk. Weighted sampling for large-scale boosting. 2008.

[14] M. Kostinger, P. Wohlhart, P. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2144–2151. IEEE, 2011.

[15] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *Computer Vision–ECCV 2012*, pages 679–692. Springer, 2012.

[16] H. Li, Z. Lin, J. Brandt, X. Shen, and G. Hua. Efficient boosted exemplar-based face detection.

[17] J. Li, T. Wang, and Y. Zhang. Face detection using surf cascade. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2183–2190. IEEE, 2011.

[18] S. Liao, A. K. Jain, and S. Z. Li. A fast and accurate unconstrained face detector. *arXiv preprint arXiv:1408.1656*, 2014.

[19] N. Markus, M. Frljak, I. S. Pandzic, J. Ahlberg, and R. Forchheimer. Object detection with pixel intensity comparisons organized in decision trees. *arXiv preprint arXiv:1305.4537*, 2014.

[20] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *ECCV*, 2014.

[21] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(1):23–38, 1998.

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, 2014.

[23] H. Schneiderman and T. Kanade. Object detection using the statistics of parts. *International Journal of Computer Vision*, 56(3):151–177, 2004.

[24] K. Sung and T. Poggio. Example-based learning for view-based human face detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(1):39–51, 1998.

[25] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[26] J. Yan, Z. Lei, L. Wen, and S. Z. Li. The fastest deformable part model for object detection.

[27] J. Yan, X. Zhang, Z. Lei, and S. Z. Li. Face detection by structural models. *Image and Vision Computing*, 2013.

[28] B. Yang, J. Yan, Z. Lei, and S. Z. Li. Aggregate channel features for multi-view face detection. *arXiv preprint arXiv:1407.4023*, 2014.

[29] C. Zhang and Z. Zhang. A survey of recent advances in face detection. Technical report, Tech. rep., Microsoft Research, 2010.

[30] L. Zhang, R. Chu, S. Xiang, S. Liao, and S. Z. Li. Face detection based on multi-block lbp representation. In *Advances in biometrics*, pages 11–18. Springer, 2007.

[31] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012.