Course name: Multimedia Processing Technique (DD026_1594)
Project number: Option 1 – 7
Project name: Survey the paper: WIDER FACE: A Face Detection Benchmark
Paper link:
https://openaccess.thecvf.com/content_cvpr_2016/papers/Yang_WIDER_FACE_A_CVPR_2016_paper.pdf

Student name: Kim woojung. 201912039
Github of project: https://github.com/Kimwoojung11/Multimedia-Processing-Technique

Report:

In this paper, it is said that face detection is the most important topic in recent years. At the same time, as many developments have been made, it is suggested that there is a slight gap between face recognition ability and what is required in reality. Therefore, to reduce this gap and facilitate future face detection research, we present a WIDER FACE dataset 10 times larger than the existing dataset. To give a brief description of the WIDER FACE dataset, it is a dataset that is 10 times larger than the existing dataset and has a rich annotation. It is also briefly introduced as an effective training source for face detection.
Given an arbitrary image, the purpose of face detection is to verify the presence of the face and, if present, return the location and range of the face. Modern face detection can easily detect the front face and is widely used.
Several available public benchmarks have contributed to face detection research, but more demanding datasets are needed as algorithms are improved and better performance is required. Currently, the face detection dataset typically contains thousands of faces, and the pose, scale, occlusion, and background clusters are limited, making it difficult to evaluate actual performance.
I will introduce the current well-known dataset that we will talk about together in the future. With AFW, FDDB, and PASCAL FACE, AFW is built using Flickrimage. It has 205 images with 473 labeled faces. For each side, the annotation includes a rectangular boundary box, six landmarks, and a pose angle. The FDDB contains annotations for 5,171 faces in 845 image sets. PASCAL FACE consists of 851 images and 1,341 annotated faces. However, due to the limited variations of

existing datasets, the performance of recent face detection algorithms is currently saturated on face detection benchmarks. For example, the highest performance in AFW is 97.2% AP, the highest recovery rate in FDDB is 91.74%, and the highest result in PASCAL FACE is 92.11% AP. So far, we have looked at the current dataset and talked about what problems there are. Now, let's look at the WIDER FACE Dataset presented in this paper. The overall contents of the WIDER FACE Dataset are as follows.

The WIDER FACE dataset is constructed based on 60 event classes. For each event class, randomly select 40%/10%/50% data as a training, validation, and test set. Two training/test scenarios are specified here. The two scenarios are as follows.

-Scenario-Ext: Face detectors are trained using external data and tested on WIDER FACE test partitions.

-Scenario-Int: Face detectors are trained using WIDER FACE training/verification partitions and tested on WIDER FACE test partitions.

Next, we will talk about how WIDER FACE Dataset collects data and specifies animation.

WIDER FACE Dataset was collected in the next three stages.

1) Event categories were defined and selected according to a large-scale multimedia ontology (LSCOM), which provides approximately 1,000 concepts related to image event analysis.

2) Images are searched using search engines such as Google and Bing. 1,000-3,000 images were collected from each category.

3) Data was organized by manually inspecting all images and filtering images without human faces. Similar images of each event category are then removed to ensure a large variety of facial features.

Annotation then labels the bounding boxes for all recognizable faces on the WIDER FACE dataset. The bounding box shall contain firmly the forehead, chin and cheek. Even if the face is covered, it is labeled with a bounding box, but there is an estimate of the size of the occlusion.

Similar to the PASCAL VOC dataset, it assigns a 'ignore' flag to all sides that are very difficult to recognize due to low resolution and small scale (below 10 pixels). After annotating the face boundary box, annotation is added to the properties of pose (normal, amorphous) and occlusion level (partial, heavy). Each annotation is

labeled by one commenter and cross-checked by two different users.
The following are attributes of the WIDER FACE Dataset. Scale, Occulation, Pose, and Event, respectively.
Scale grouped faces on three scales, small (between 10-50 pixels), medium (between 50-300 pixels), and large (over 300 pixels), according to image size (in pixels).
Large and medium-sized scales achieved a high detection rate (over 90%) with 8,000 per image. In the case of small-scale, the detection rate continues to be less than 30%, even though the number of proposals has been increased to 10,000.
Occulsion is an important factor in evaluating facial detection performance. Similar to recent research, we treat Occulsion as an attribute and assign faces to three categories: (no occlusion, partial occlusion, heavy occlusion).
Specifically, we choose 10% of faces as an example. Each exemplary side is annotated with two bounding boxes representing the 'visible face range' and the 'full face range'. We calculate the occlusion ratio at 1 divided by the visible face area by the total face area
The face is defined as "partially covered" if 1% to 30% of the total face area is covered. Faces with more than 30% of the occlusion area are classified as "severe obstruction."
It shows that the detection rate decreases as the level of occlusion increases. The detection rate of partial or severely obstructed faces is less than 50% out of 8,000 proposals.
Pose defines two pose deformation levels: typical and unstructured. The face is represented by a typical annotation under two conditions. Faces with typical poses are much more difficult to detect.
Event characterizes each event with three components: scale, occlusion, and pose to evaluate the effect of the event on face detection. For each factor, the detection rate for a particular event class is calculated, and then the detection rates are ranked in ascending order. Depending on the ranking, the event is divided into three divisions: Easy (41-60 class), Medium (21-40 class), and Hard (1-20 class).
Subsequently, this paper attempts to establish a robust criterion for the WIDER

FACE dataset.
To address the high variability of scale, we propose a multi-stage two-stage cascade framework and use segmentation and conquest strategies.
Specifically, we train a set of face detectors, each of which deals with a relatively small range of faces. The face detector consists of two stages.
The first step produces a multi-scale proposal in a fully convolutional network.
The second step is a multi-task convolutional network that generates face and non-face-to-face predictions of candidate windows obtained in the first step and simultaneously predicts face positions.
The multi-scale proposal here is to jointly train a series of fully convolutional networks for face classification and scale classification. Groups faces into four categories according to image size. For each group, divide it further into three subclasses. Each network is trained with an image patch with a size of an upper limit scale. Align the faces at the center of the image patch with positive samples and assign scale class labels based on each group's predefined scale subclass.
For face classification, a positive label is assigned if the IoU between the proposed window and the survey information boundary box is greater than 0.5, otherwise it is a negative window.
For boundary box regression, each proposal predicts the location of the nearest ground truth boundary box. If the proposed window is false positive, the CNN outputs a vector of [-1, -1, -1, -1]. We adopt Euclidean loss and cross entropy loss for boundary box regression and face classification, respectively.
Compare benchmarks and WIDER FACE results and talk about the results of the experiment. Among the benchmarks, there are four main Face detection algorithms: VJ, ACF, DPM, and Face, but here we will only talk about ACF and Face, which will be compared to WIDER FACE. In addition, the PASCAL VOC evaluation metric was used for evaluation.
In benchmarks, ACF and Faculty showed relatively superior performance compared to the other two algorithms. So, comparing these two algorithms with WIDER FACE, the model retrained to WIDER FACE was improved by 5.4% and 4.2% respectively compared to the reference models. This demonstrates the effectiveness of the WIDER FACE dataset as a training source.
In this paper, we propose a large annotated WIDER FACE dataset to train and

evaluate facial detection algorithms. We benchmark four typical face detection methods. Even considering easy subsets (typically subsets with a face height of more than 50 pixels), existing state-of-the-art algorithms reach approximately 70% APs. The face captured by surveillance cameras in public places or events is typically small, covered, and typically posed. These faces undoubtedly suggest the most interesting yet decisive thing to detect for further investigation.