Gaussian Vector: An Efficient Solution for Facial Landmark Detection

Yilin Xiong[1,2][0000−0003−1989−916X], Zijian Zhou[2][0000−0003−3315−3962], Yuhao Dou[2][0000−0002−4407−9609], and Zhizhong Su[2][0000−0003−2312−9985]

[1] Central South University, Changsha, Hunan, CN
[2] Horizon Robotics
yilin.xiong@csu.edu.cn
{zijian.zhou, yuhao.dou, zhizhong.su}@horizon.ai

1 supplementary material

1.1 About FR@(10%) in Table 3

In Table 3, we achieve better NME and AUC than AWing[42], but worse FR. As we presented in Sec 4.2, NME calculates the overall mean error of the test set, and AUC evaluates the samples with error lower than threshold. But FR is the proportion of samples with error higher than threshold in the test set. The results show that, compared with AWing, our approach is better overall but worse in extreme hard samples. (e.g. heavy occlusion). To achieve SOTA, AWing used boundary map and multi-stage supervision, resulting in increased network complexity. As first introduced in the LAB [43], the boundary map provides better face structure information to deal with hard samples. However, it is complicated to generate and not suitable for sparse landmarks (e.g. 5 points). To prove our vector label works, we try to keep our network simple enough. Even so, our method goes beyond AWing on the NME and AUC.

1.2 Analysis

We can see that our proposed method surpasses the previous heatmap based methods by a large margin. **How do the vector supervision and BPM contribute to model performance?** Here we analyze the merits in three aspects.

Firstly, vector supervision alleivates the imbalance of foreground and background to a great extent. For example, in the typical Gaussian heatmap with a shape of 64×64, standard deviation σ=2, the foreground pixels take up only 4%. The extreme ratio leads the foreground to a less important position. In the early training stage, the optimizer makes great effort to push all pixels to zero values, which slows down the convergence. Even after a period of training, the accumulation of small errors on the background still affects the training optimization, which is meaningless for landmark detection. The proportion rises to 20% when vector supervision is used. Consequently, the whole training process becomes more efficient.

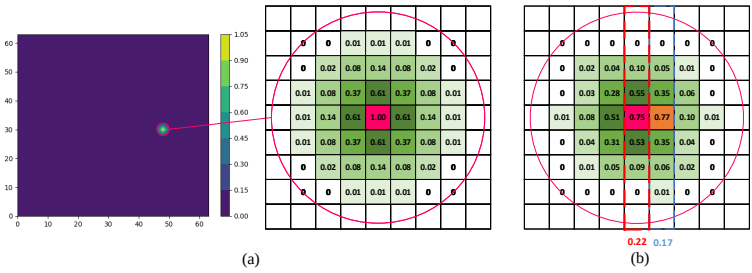

Fig. 1. (a) The ground truth Gaussian heatmap. (b) A tiny shift on the pixel with the highest response causes location error in predicted heatmap. The ground truth location (*red* pixel with 0.75) is smaller than the one on the right (*orange* pixel with 0.77), bringing about extra location error. With l=1, however, the mean of corresponding two columns is 0.22 (*red* dashed region) and 0.17 (*blue* dashed region). The vertical band pooling helps eliminate the error by average the pixels in the band.

Secondly, in the heatmap based methods, only the pixels in the neighbor domain of the maximum one are used to get the final landmark coordinates. Most of the spatial information is discarded even though we devote much energy to optimize it. Worse yet, a tiny shift of the maximum pixel on the heatmap may increase location error. Fig. 1(a) shows heatmap supervision and Fig. 1(b) shows maximum pixel shifting on the heatmap prediction incurs an evident error. Overwhelming the heatmap based methods, the proposed Band Pooling Module comprehensively makes use of the spatial information. It demonstrates that BPM stengthens the resistibility of the maximum pixel shifting and leads to more robust prediction.

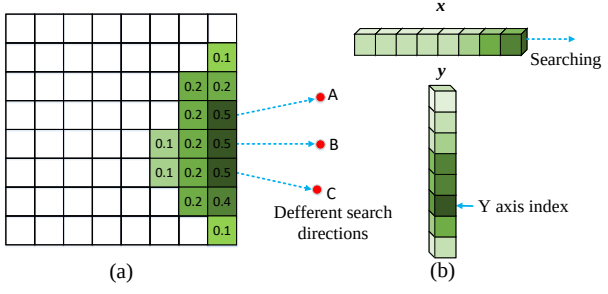

Fig. 2. (a) The maximum pixel locates on the right edge of the heatmap. Regression error caused by weak supervision usually brings about multiple maximum pixel or maximum pixel shifting. Therefore, it is difficult to locate the correct starting point and search direction. (b) Vector supervision decouples X and Y axes, as a result, we are able to locate the correct Y axis index from vector $\mathbf{y}$ and search along the right direction.

Finally, Beyond Box Strategy helps our model look outside and predict the landmarks out of the bounding box. With this strategy, we search the peak of quasi-Gaussian distribution towards a single direction. However, it is too hard to assume an exact 2D distribution in a similar way for heatmap based methods, because we are usually not sure about the starting point and search direction. As shown in Fig. 2(a), multiple maximum pixels locate on the right edge of the heatmap. As a result, we cannot determine which point to start with and which direction to search towards. Since we convert the heamtp into a pair of vectors like Fig. 2(b), the two axes are decorrelated naturally. We can easily get the accurate starting point from the maximum of vector $\mathbf{y}$, and search on vector $\mathbf{x}$ along the right side. Therefore, even if the landmarks fall outside, Beyond Box Strategy is able to settle this trouble in most cases.