Quality Aware Network for Set to Set Recognition

Yu Liu, Junjie Yan, Wanli Ouyang

SenseTime Group Limited The University of Sydney

SENSETIME
Focused on innovative computer vision and deep learning technologies

THE UNIVERSITY OF SYDNEY

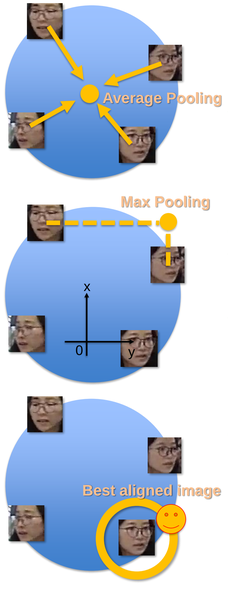
Our code on GitHub

IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 2017 HONOLULU

Set to set recognition

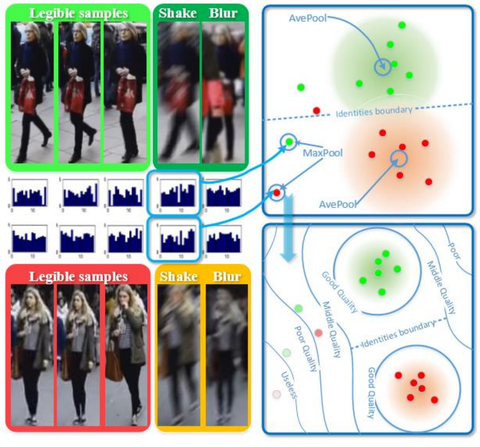
DIFFERENT FRAMES HOLD DIFFERENT QUALITIES
BUT COMPLEMENTARY INFORMATION

Pooling all frames for set level representations



STRAIGHT FORWARD METHODS

POOLING BY 'QUALITY'

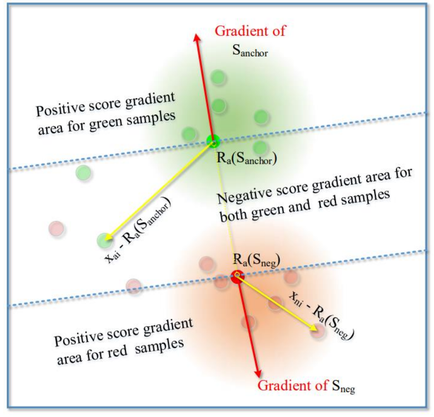


QAN WEAKEN THE NOISY SAMPLES AND NARROW DOWN IDENTITIES' VARIANCES.

Learning Quality without label

WEIGHTED POOLING MECHANISM

$$\mathcal{F}(R_{I_1}, R_{I_2}, \cdots, R_{I_N}) = \frac{\sum_{i=1}^{N} \mu_i R_{I_i}}{\sum_{i=1}^{N} \mu_i} \qquad \mu_i = Q(I_i)$$



AUTOMATIC QUALITY LEARNING

$$\frac{\partial \mathcal{F}}{\partial R_{I_i}} = \frac{\partial R_a(S)}{\partial R_{I_i}} = \mu_i$$

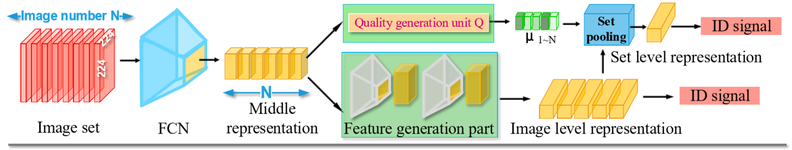
$$\frac{\partial \mathcal{F}}{\partial \mu_i} = \frac{\partial R_a(S)}{\partial \mu_i} = R_{I_i} - R_a(S)$$

$$\frac{\partial L_{veri}}{\partial R_{I_i}} = \frac{\partial R_a(S)}{\partial R_{I_i}} \cdot \frac{\partial L_{veri}}{\partial R_a(S)} = \frac{\partial L_{veri}}{\partial R_a(S)} \cdot \mu_i$$

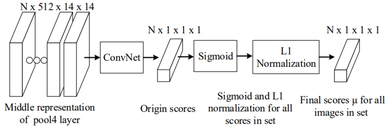
$$\frac{\partial L_{veri}}{\partial \mu_i} = \frac{\partial R_a(S)}{\partial \mu_i} \cdot \left(\frac{\partial L_{veri}}{\partial R_a(S)}\right)^T = \sum_{j=1}^{D} \left(\frac{\partial L_{veri}}{\partial R_a(S)_j} \cdot (x_{ij} - R_a(S)_j)\right)$$

Network structure

WE IMPLEMENT THIS MECHANISM IN AN END-TO-END CNN



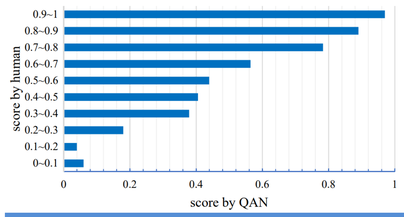
DETAILS IN SCORE GENERATION UNIT Q



Quality learned by QAN

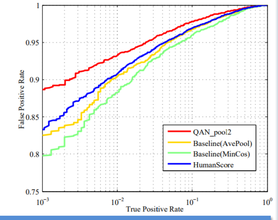
IMAGES WITH THEIR QUALITIES

QUALITY BY QAN IS SIMILAR WITH BUT BUTTER THAN THAT BY HUMAN



COMPARISON WITH HUMAN LABELED SCORE

ROC ON YOUTUBE FACE DATASET



Enhancing Acc. on Face and Human recognition

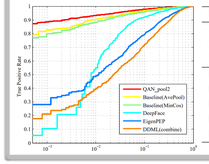
PRID2011

Methods	CMC1	CMC5	CMC10	CMC20
QAN	**90.3**	**98.2**	**99.32**	**100.0**
CNN+AvePool	81.3	96.6	98.5	99.6
CNN+Min(cos)	69.8	91.3	97.1	99.8
CNN+RNN [36]	70	90	95	97
STFV3D [22]	42.1	71.9	84.4	91.6
TDL [40]	56.7	80.0	87.6	93.6
eSDC [34]	48.3	74.9	87.3	94.4
DVR [34]	40.0	71.7	84.5	92.2
LFDA [25]	43.7	72.8	81.7	90.9
KISSME [16]	34.4	61.7	72.1	81.0
LADF [21]	47.3	75.5	82.7	91.1
TopRank [19]	31.7	62.2	75.3	89.4

iLIDS-VID

Methods	CMC1	CMC5	CMC10	CMC20
QAN	**68.0**	**86.8**	95.4	97.4
CNN+AvePool	60.6	84.9	89.8	93.6
CNN+Min(cos)	49.3	79.4	88.2	91.9
CNN+RNN [36]	58	84	91	96
STFV3D [22]	37.0	64.3	77.0	86.9
TDL [40]	56.3	87.6	**95.6**	**98.3**
eSDC [34]	41.3	63.5	72.7	83.1
DVR [34]	39.5	61.1	71.7	81.0
LFDA [25]	32.9	68.5	82.2	92.6
KISSME [16]	36.5	67.8	78.8	87.1
LADF [21]	39.0	76.8	89.0	96.8
TopRank [19]	22.5	56.1	72.7	85.9

YOUTUBE FACE DATASET



Method	Accuracy(%)	AUC
QAN	**96.17± 0.09%**	**99.14± 0.12%**
CNN+AvePool	95.46± 0.07%	98.66± 0.04%
CNN+Min(cos)	94.87± 0.10%	98.37± 0.06%
NAN [38]	95.52±0.06%	98.7%
FaceNet [26]	95.12±0.39%	-
DeepID2+ [29]	93.2±0.2%	-
DeepFace-single [30]	91.4±1.1%	96.3%
EigenPEP [18]	84.8±1.4%	92.6%

PERSON RE-IDENTIFICATION

I-JBA

TPR@FPR	1e-3	1e-2	1e-1
QAN	**89.31±3.92%**	**94.20±1.53%**	**98.02±0.55%**
CNN+AvePool	85.30±3.48%	93.81±1.44	97.85±0.61%
CNN+Min(cos)	82.74±3.61%	92.06±1.98	97.29±0.67%
NAN [38]	78.5±2.8%	89.7±1.0%	95.9±0.5%
DCNN+metric [4]	-	78.7±4.3%	94.7±1.1%
LSFS [31]	51.4±6.0%	73.3±3.4%	89.5±1.3%
OpenBR [15]	10.4±1.4%	23.6±0.9%	43.3±0.6%