CAN HOLISTIC REPRESENTATIONS BE USED FOR FACE BIOMETRIC QUALITY ASSESSMENT?

Samarth Bharadwaj, Mayank Vatsa and Richa Singh

IIIT-Delhi, India

ABSTRACT

A face quality metric must quantitatively measure the *usability* of an image as a biometric sample. Though it is well established that quality measures are an integral part of robust face recognition systems, automatic measurement of *biometric quality* in face is still challenging. Inspired by scene recognition research, this paper investigates the use of holistic super-ordinate representations, namely, *Gist* and sparsely pooled Histogram of Orientated Gradient (HOG), in classifying images into different quality categories that are derived from matching performance. The experiments on the CAS-PEAL and SCFace databases containing covariates such as illumination, expression, pose, low-resolution and occlusion by accessories, suggest that the proposed algorithm can efficiently classify input face image into relevant quality categories and be utilized in face recognition systems.

Index Terms— biometrics, face quality assessment, performance prediction.

1. INTRODUCTION

Biometric systems deployed in unconstrained environments, for example, large-scale identity projects such as Aadhaar and US-Visit, encounter varying quality of input face samples as shown in Fig. 1. To improve the performance, usability, and robustness of such systems, recent research in face biometrics use *quality* of the sample not only to reject the poorly captured samples but also within the recognition process. The active involvement of quality scores beyond the capture stage encourages the formulation of more complex and accurate quality assessment techniques. Current research in face recognition generally use simple image processing algorithms that are able to assess image degradations due to noise, compression or illumination. While the quality of a face image is susceptible to degradation during capture and storage, it may also have low quality by its very nature. For example, a high resolution face image with acute pose is of low biometric quality, irrespective of the high image quality. The complexity of the problem is further exacerbated by the lack of consensus in literature on facial (biometric) features.

{samarthb, mayank, rsingh}@iiitd.ac.in

Fig. 1. Face images of varying quality encountered by a face recognition systems.

Table 1 summarizes important approaches in face quality assessment. Early research in face quality [1, 2] focuses on complete automation of essential capture guidelines in face standards such as ICAO and ISO/JEC $19794-5$ [3]. The effects of resolution and capture conditions, with an analysis of *subjective* and *objective* covariates of face biometric in Face Recognition Vendor Test (FRVT) 2006 is presented in [4, 5]. A probabilistic approach for performance prediction using background information (capture conditions, gender, race) is discussed [6]. Further, considerable research on leveraging simple quality metrics to improve multibiometrics recognition is summarized in [7]. Image quality assessment metrics, with focus on perceptual quality, are reviewed in [8].

This research explores a new direction in *face quality assessment using holistic representations*. Scene recognition techniques extensively use holistic features that are designed to encode commonalities in a large collection of scene images for image classification. As opposed to biometric features, that encode *unique* attributes of an image, scene recognition techniques effectively encode abstract and categorical features of an image such as vertical or horizontal structures in city images, and openness in landscape images. It is our assertion that these features can be used to *predict* the *usability* of a face image and segregate face images into abstract categories that are indicative of quality. The experiments on a heterogenous database consisting of several covariates show that holistic image descriptors are able to successfully categorize biometric images (using a classifier) into quality bins ranging from *poor* to *excellent* quality, *that correlates with recognition performance*. Further, as a case study, improved face recognition performance is observed when the proposed approach is used to reject poor quality samples.

Table 1. Summary of a representative list of existing approach in face quality.

Technique	Description
Subasic *et al.* [1]	17 automatic tests for $ICAO$ standards.
Hsu *et al.* [2]	Automatic evaluator of ISO/JEC19794-5 face standards.
Youmaran and Adler [9]	Biometric information defined from information theory.
Gao *et al.* [10]	Asymmetry in LBP features as a measure of the quality.
Zhang *et al.* [11]	Asymmetry using SIFT features.
Wong *et al.* [12]	Comparison of a facial image with *ideal* face models.
Nasrollahi *et al.* [13]	Geometrical pose estimation using face bounding box.
Yao *et al.* [14]	Sharpness measure for frame selection.
Proposed	**Use holistic descriptors with match score based pseudo-labels for quality prediction.**

2. QUALITY ASSESSMENT OF FACE BIOMETRIC

Research in scene recognition has shown that holistic representation of an image is consistent in abstractly classifying images into broad categories such as *buildings, coastline* and *forests*. Inspired by this observation, a learning based approach to quality assessment is proposed in this research. The mapping between recognition performance based *quality labels* and holistic representation of images is learned in a supervised setting and utilized to predict quality.

First, quality labels are generated based on match score distribution obtained from a powerful matcher. Next, these labels are assigned to a set of training images with different image and biometric degradations (illumination, low resolution, occlusion, and expression). A non-linear relation between these labels and multi-dimensional holistic descriptors is learned using a multi-class classifier.

2.1. Face Quality as Match Score Predictor

We develop the intuition of such an approach to quality assessment, from quality based match score prediction. As shown by Grother and Tabassi [15], there is a relationship between quality of a biometric sample and recognition accuracy. For a quality assessment algorithm (Q_1) that produces a scalar quality metric q, the match score $s_{d,d'}$ between the samples d and d' can be modeled using a predictor P as,

$$s_{d,d'} = P(Q_1(d), Q_1(d')) + \epsilon_{d,d'}, \qquad (1)$$

The predictor P estimates the similarity score based on the quality of the templates ($\epsilon_{d,d'}$ is the error in that prediction). This problem of biometric match score prediction is challenging since Q produces a single quality value. However, Vatsa *et al.* [16] and Bhatt *et al.* [17] present evidence indicating that a comprehensive quality measure must be a vector rather than a scalar. Hence, Eq. 1 is redefined as,

$$s_{d,d'} = P(\vec{q} = Q_2(d), \vec{q'} = Q_2(d')) + \epsilon_{d,d'}, \qquad (2)$$

where $\vec{q}$ and $\vec{q'}$ are the quality vectors of samples d and d' that may provide more information for P. In this research, $\vec{q}$ is a multi-dimensional holistic representation of the probe image that preserves non-localized, categorical information of the image.

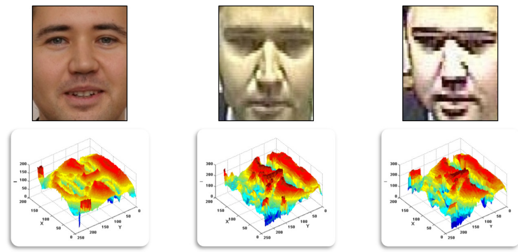


Fig. 2. Face images with degradations exhibit more *roughness*, evident from the surface plots (z-axis is pixel intensity (I)). Roughness can be captured with holistic features and may be indicative of biometric quality.

2.2. Holistic Image Representations

In this research, two prominent holistic representations, Gist [18] and sparsely pooled HOG [19] are considered. As illustrated in Fig. 2, poor quality face images have a typical *roughness* in intensity values as compared to a good quality image. The abstract and non-localized nature of Gist and HOG make them good candidates to assess biometric quality. Further, compared to local image descriptors, the feature length of Gist (512) and HOG (81) can ensure low computational time for quality assessment. A brief summary of Gist and HOG is presented below.

Gist: Olivia and Torralba [18] propose a holistic representation of the spatial envelope of a scene image. Rather than viewing an image as a configuration of objects, a unitary model is used. The spatial properties of the image are well preserved in such a representation, referred to as *Gist*. A set of five perceptual dimensions, namely, naturalness, openness, roughness, expansion and ruggedness are used to compute low dimensional, holistic representation of the image. These coarse features are highly abstract and obtained from the amplitude spectrum of the windowed Fourier transform. These perceptual properties are correlated with the second-order statistics and spatial arrangement of structured components in the image.

HOG: Dalal and Triggs [19] present a simple descriptor known as histogram of orientated gradient that is popularly used for humans, vehicles and animals detection in still imagery and videos due to the low computation time as well as high accuracy. Unidirectional gradient kernel is applied on a normalized image to obtain the orientations in local regions. These local descriptors are spatially pooled to obtain a *holistic representation* of each region (3×3 blocks). Further, the histograms with 9 bins are normalized by $k{-}norm$ operation.

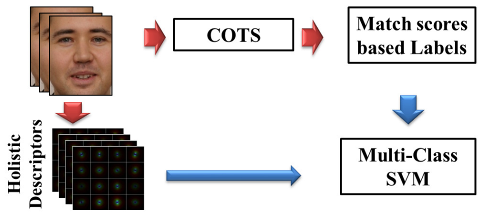


Fig. 3. The training process of the proposed approach.

2.3. Quality Labels based on Face Matcher

As shown in Fig. 3, the relationship between an image representation and quality label is learned from a training set using a non-linear classifier. The training samples are annotated based on the identification performance on the training set, inspired from [15]. The steps to obtain the quality label are as follows:

- A matching algorithm is used to obtain the match scores (s) on a training data that consists of a good quality (studio quality) image and several probe images of varying quality per subject. In order to minimize the misclassification rate, match scores obtained from two commercial off-the-shelf (COTS) face recognition system are fused using sum rule.
- The genuine match scores ($s_{d=d'}$) are Z-normalized and segregated into two sets. *Correct matches* refer to those genuine scores that result in Rank-1 matching. The remaining are referred to as *Incorrect matches*. Next, the empirical cumulative distribution function (ECDF) of both the sets are obtained (cdf_C,cdf_I), as illustrated in Fig. 4. Further, the training probe samples are labeled, as *excellent*, *good*, *fair* and *poor* corresponding to the bins of match scores. The bin thresholds and number of bins can be varied according to the specific application scenarios.
- A one-vs-all multi-class SVM is trained for four bins of quality with the holistic descriptor as the input feature. The label corresponding to the most confident positive classification of SVM is selected in the testing phase.

Table 2. Summary of databases.

Database	Subjects (Train/Test)	Description
SCFace [20]	130 (39/91)	pose, low resolution
CAS-PEAL [21]	1040 (312/728)	pose, illumination, expression, accessories, background, distance
Combined	1170 (351/819)	all of the above

3. EXPERIMENT AND ANALYSIS

A face quality assessment technique must be aware of all the degradations that are encountered in face modality. A single

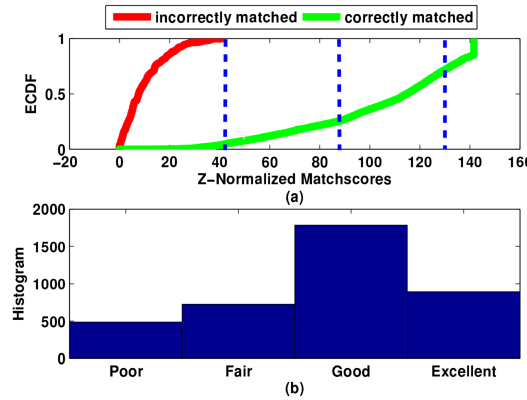


Fig. 4. (a) The empirical cumulative density function (ECDF) of the z-normalized match scores (-20 to 160), (b) the number of samples per quality bin obtained for training.

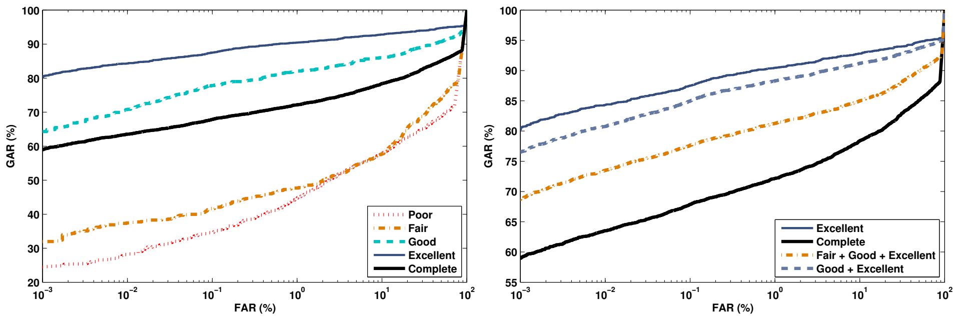
Table 3. Performance of COTS on each quality bin.

	Quality Bin	Count	% Hist. Overlap	Rank-1 %	EER
HOG	Excellent	390	2.65%	89.48%	7.58%
	Good	278	7.27%	75.89%	12.84%
	Fair	4198	15.06%	74.65%	20.15%
	Poor	639	16.18%	48.98%	21.50%
	Good + Excellent	668	17.43%	83.83%	10.03%
	Fair + Good + Excellent	4866	25.63%	75.91%	18.72%
Gist	Excellent	871	4.97%	91.10%	7.48%
	Good	2766	8.23%	82.43%	13.55%
	Fair	713	25.59%	45.86%	30.86%
	Poor	1155	32.43%	38.26%	34.11%
	Good + Excellent	3637	12.74%	89.03%	8.91%
	Fair + Good + Excellent	4350	19.68%	81.95%	14.28%
	Complete	5505	28.46%	72.78%	19.51%

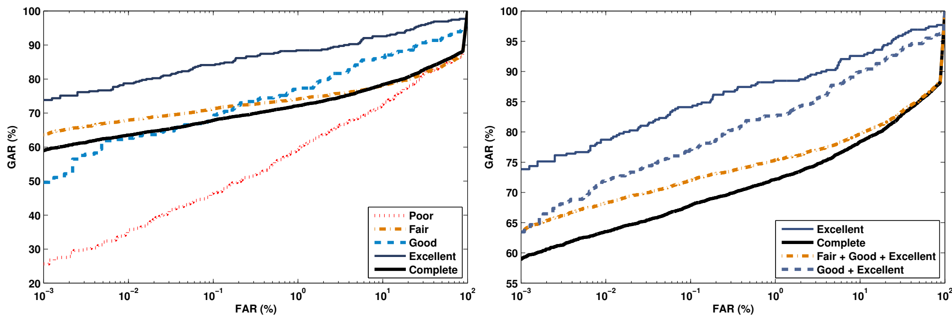
face database is usually collected in similar settings and may lead to bias in quality assessment approaches that are based on training. Hence, in this research, a heterogeneous combination of two face databases, namely, the SCFace [20] and CAS-PEAL [21], with pose, illumination, expression, accessories, background, distance and resolution variations is used. Images corresponding to 30% of the subjects are used as training and the remaining as testing (summarized in Table 2). In both the training and testing phases, a single good quality image is used as gallery and the remaining images are used as probe. All the training samples corresponding to quality bins are used to train SVM and the parameters are obtained via grid search, with radial basis function as the kernel. To evaluate the correctness of quality labels, the identification and verification performance of each bin are computed separately using the better performing COTS, similar to the experimental procedure in [15].

- On the training database, the fusion of two COTS yields the rank-1 identification accuracy of 91.69%. Hence, all the *Incorrect matches* are marked as *poor* quality ($cdf_I^{-1}(1)$[1]). Further, $cdf_I^{-1}(1)$ to $cdf_C^{-1}(0.25)$ are la-

[1]Here, $cdf_X^{-1}(a)$ corresponds to the value of the random variable X



(a) GIST



(b) HOG

Fig. 5. Verification performance of testing images when segregated into quality bins (left) and when lower quality bins are discarded (right) using a) Gist and b) HOG.

beled *fair* quality, $cdf_C^{-1}(0.25)$ to $cdf_C^{-1}(0.75)$ as *good* and beyond $cdf_C^{-1}(0.75)$ as *excellent*. As mentioned, this configuration may be application dependent.

- Table 3 and Fig. 5 show the performance of COTS on each of the quality bins obtained from both GIST and HOG. Better performance is observed for quality bins classified as *excellent* and *good* compared to *fair* and *poor*. Further, the percentage overlap for the genuine and imposter distributions is also increased for lower quality images. The difference in performance of each bin indicates the validity of the assigned bin labels.
- In several applications of biometrics such as Aadhaar and US-Visit, low quality image samples are rejected to maintain the integrity of the database and to ensure high recognition accuracy. The proposed algorithm can be utilized to reject low quality samples. Fig. 5 and Table 3 show improved performance compared to the complete database, when images classified as *poor* and/or *fair* are removed, indicating a direct relationship of the proposed metric with system performance.
- Fig. 6 illustrates samples from the database classified into a particular quality bin. The illustrated instances are obtained from the set of images classified to a quality bin by both Gist and HOG. It can be observed that the classification correlates well with visual inspection.

where the cumulative density is less than or equal to a.

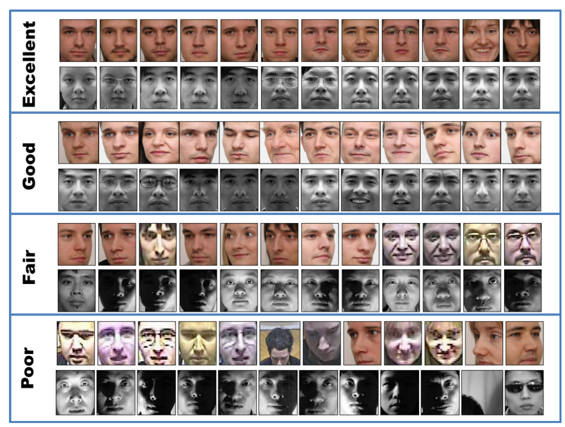


Fig. 6. Sample images of four quality bins obtained from the proposed approach (common to both Gist and HOG).

4. CONCLUSION

Quality metrics are an important ingredient to improve the robustness of large scale real-world face biometric systems. This research investigates the possibility of using holistic representation of an image for quality assessment. The results with Gist and HOG show promise towards a robust solution to the important problem of quality assessment in face biometrics. By further evaluating the effects of each quality class on recognition accuracy, the techniques described in this research can also be used for classifier performance prediction.

5. REFERENCES

[1] M. Subasic, S. Loncaric, T. Petkovic, H. Bogunovic, and V. Krivec, "Face image validation system," in *Proceedings of International Symposium on Image and Signal Processing and Analysis*, 2005, pp. 30–33.

[2] R.L.V. Hsu, J. Shah, and B. Martin, "Quality assessment of facial images," in *Proceedings of Biometric Consortium Conference*, 2006, pp. 1–6.

[3] ISO/IEC, "ISO/IEC 19794 Biometric Data Interchange Formats," *ISO/IECJTC1/SC37/ Working Group 3*, 2005.

[4] J. R. Beveridge, G.H. Givens, P. J. Phillips, B.A. Draper, D.S. Bolme, and Y.M. Lui, "Focus on quality, predicting FRVT 2006 performance," in *Proceedings of IEEE International Conference on Automatic Face Gesture Recognition*, 2008, pp. 1–8.

[5] J. R. Beveridge, P. J. Phillips, G. H. Givens, B. A. Draper, M. N. Teli, and D. S. Bolme, "When high-quality face images match poorly," in *Proceedings of IEEE international Conference on Automatic Face Gesture Recognition and Workshops*, 2011, pp. 572 –578.

[6] J. R. Beveridge, G. H. Givens, P. J. Phillips, B. A. Draper, D. S. Bolme, and Y. M. Lui, "FRVT 2006: Quo vadis face quality," *Image and Vision Computing*, vol. 28, no. 5, pp. 732–743, 2010.

[7] N. Poh and J. Kittler, "A unified framework for multimodal biometric fusion incorporating quality measures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 3–18, 2011.

[8] Z. Wang and A.C. Bovik, *Modern Image Quality Assessment*, vol. 2, Synthesis Lectures on Image, Video, and Multimedia Processing, 2006.

[9] R. Youmaran and A. Adler, "Measuring biometric sample quality in terms of biometric information," in *Proceedings of Biometric Consortium*, 2006, pp. 1–6.

[10] X. Gao, S. Li, R. Liu, and P. Zhang, "Standardization of face image sample quality," in *Proceedings of International Conference on Biometrics*, 2007, vol. 4642, pp. 242–251.

[11] G. Zhang and Y. Wang, "Asymmetry-based quality assessment of face images," in *Proceedings of International Symposium on Visual Computing*, 2009, vol. 5876, pp. 499–508.

[12] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B.C. Lovell, "Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, 2011, pp. 74–81.

[13] K. Nasrollahi and T. Moeslund, "Face quality assessment system in video sequences," in *Biometrics and Identity Management*, 2008, vol. 5372, pp. 10–18.

[14] Y. Yao, B. Abidi, and M. Abidi, "Quality assessment and restoration of face images in long range/high zoom video," in *Face Biometrics for Personal Identification*, 2007, pp. 43–60.

[15] G. Patrick and E. Tabassi, "Performance of biometric quality measures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 531–524, 2007.

[16] M. Vatsa, R. Singh, A. Noore, and A. Ross, "On the dynamic selection of biometric fusion algorithms," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 3, pp. 470–479, 2010.

[17] H. Bhatt, S. Bharadwaj, M. Vatsa, R. Singh, A. Ross, and A. Noore, "A framework of quality-based biometric classifier selection," in *Proceedings of IEEE/IAPR International Joint Conference on Biometrics*, 2011, pp. 1–7.

[18] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of IEEE international Conference on Computer Vision and Pattern Recognition*, 2005, vol. 1, pp. 886–893.

[20] M. Grgic, K. Delac, and S. Grgic, "SCface surveillance cameras face database," *Multimedia Tools and Applications*, vol. 51, no. 3, pp. 863–879, 2011.

[21] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao, "The CAS-PEAL large-scale chinese face database and baseline evaluations," *IEEE Transactions on System Man, and Cybernetics-A*, vol. 38, no. 1, pp. 149–161, 2008.