A New Look at Filtering Techniques for Illumination Invariance in Automatic Face Recognition

Ognjen Arandjelović Roberto Cipolla
Department of Engineering
University of Cambridge
Cambridge, UK CB2 1PZ
{oa214,cipolla}@eng.cam.ac.uk

Abstract

Illumination invariance remains the most researched, yet the most challenging aspect of automatic face recognition. In this paper we propose a novel, general recognition framework for efficient matching of individual face images, sets or sequences. The framework is based on simple image processing filters that compete with unprocessed greyscale input to yield a single matching score between individuals. It is shown how the discrepancy between illumination conditions between novel input and the training data set can be estimated and used to weigh the contribution of two competing representations. We describe an extensive empirical evaluation of the proposed method on 171 individuals and over 1300 video sequences with extreme illumination, pose and head motion variation. On this challenging data set our algorithm consistently demonstrated a dramatic performance improvement over traditional filtering approaches. We demonstrate a reduction of 50–75% in recognition error rates, the best performing method-filter combination correctly recognizing 96% of the individuals.

1. Introduction

In this work we are interested in illumination invariance for automatic face recognition (AFR), and, in particular, the case when both training and novel data to be matched are image *sets or sequences*. Invariance to changing lighting is perhaps the most significant practical challenge for AFR. The illumination setup in which recognition is performed is in most cases impractical to control, its physics difficult to accurately model and face appearance differences due to changing illumination are often larger than differences between individuals [1]. Additionally, the nature of most real-world AFR application is such that prompt, often real-time system response is needed, demanding appropriately efficient matching algorithms.

In this paper we propose a novel framework for rapid recognition under varying illumination, based on simple image filtering techniques. The framework is very general: we demonstrate that it offers a dramatic performance improvement to a wide range of filters and different baseline matching algorithms, without sacrificing their online efficiency.

1.1. Previous Work – AFR across Illumination

Two of the most influential approaches to achieving robustness to changing lighting conditions are the illumination cones of Belhumeur *et al.* [4, 10] and the 3D morphable model of Blanz and Vetter [5]. In [4] the authors showed that the set of images of a convex, Lambertian object, illuminated by an arbitrary number of point light sources at infinity, forms a convex polyhedral cone in the image space with dimension equal to the number of distinct surface normals. In [10], Georghiades et al. successfully used this result for AFR by reilluminating images of frontal faces. In the 3D morphable model method, parameters of a complex generative model which includes the pose, shape and albedo of a face are recovered in an analysis-by-synthesis fashion.

Both illumination cones and the 3D morphable model have significant shortcomings for practical AFR use. The former approach assumes very accurately registered face images, illuminated from seven to nine different well-posed directions for each head pose. This is difficult to achieve in practical imaging conditions (see §3 for typical image data quality). On the other hand, the 3D morphable model requires a (in our case prohibitively) high resolution [7], struggles with non-Lambertian effects and multiple light sources, has convergence problems in the presence of background clutter and partial occlusion (glasses, facial hair), and is very computationally demanding.

Most relevant to the material presented in this paper are methods that can be broadly described as quasi illumination-invariant *image filters*. These include high-pass [3] and locally-scaled high-pass filters [17], directional

derivatives [1, 7] and edge-maps [1], to name a few. These are most commonly based on very simple image formation models, for example modelling illumination as a spatially low-frequency band of the Fourier spectrum and identity-based information as high-frequency [3, 8]. Methods of this group can be applied in a straightforward manner to either single or multiple-image AFR and are often extremely efficient. However, due to the simplistic nature of the underlying models, in general they do not perform well in the presence of extreme illumination changes.

2. Adapting to Data Acquisition Conditions

The framework proposed in this paper is most closely motivated by the findings first reported in [2]. In that work several AFR algorithms were evaluated on a large database using (i) raw greyscale input, (ii) a high-pass (HP) filter and (iii) the Self-Quotient Image (QI) [17]. Both the high-pass and even further Self Quotient Image representations produced an improvement in recognition for all methods over raw grayscale, which is consistent with previous findings in the literature [1, 3, 8, 17]. Of importance to this paper is that it was also examined in which cases these filters help and how much depending on the data acquisition conditions. It was found, consistently over different algorithms, that recognition rates using greyscale and either the HP or the QI filter negatively correlated (with $\rho \approx -0.7$).

This is an interesting result: it means that while on average both representations increase the recognition rate, they actually *worsen* it in "easy" recognition conditions when no normalization is needed. The observed phenomenon is well understood in the context of energy of intrinsic and extrinsic image differences and noise (see [18] for a thorough discussion). Higher than average recognition rates for raw input correspond to small changes in imaging conditions between training and test, and hence lower energy of extrinsic variation. In this case, the two filters decrease the SNR, worsening the performance. On the other hand, when the imaging conditions between training and test are very different, normalization of extrinsic variation is the dominant factor and performance is improved.

This is an important observation: it suggests that the performance of a method that uses either of the representations can be increased further by detecting the difficulty of recognition conditions. In this paper we propose a novel learning framework to do this.

2.1. Adaptive Framework

Our goal is to implicitly learn how similar the novel and training (or *gallery*) illumination conditions are, to appropriately emphasize either the raw input guided face comparisons or of its filtered output. Fig. 1 shows the difficulty

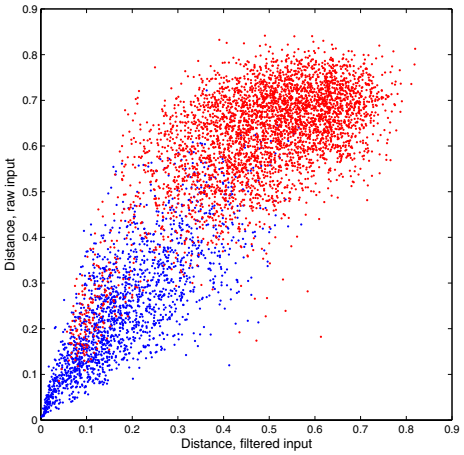


Figure 1. *Distances* $(0-1)$ *between sets of faces – interpersonal and intrapersonal comparisons are shown respectively as large red and small blue dots. Individuals are poorly separated.*

of this task: different classes (i.e. persons) are not well separated in the space of 2D feature vectors obtained by stacking raw and filtered similarity scores.

Let $\{\mathcal{X}_1, \ldots, \mathcal{X}_N\}$ be a database of known individuals, $\mathcal{X}$ novel input corresponding to one of the gallery classes and $\rho()$ and $F()$, respectively, a given similarity function and a quasi illumination-invariant filter. We then express the degree of belief η that two face sets $\mathcal{X}$ and $\mathcal{X}_i$ belong to the same person as a weighted combination of similarities between the corresponding unprocessed and filtered image sets:

$$\eta = (1-\alpha^*)\rho(\mathcal{X}, \mathcal{X}_i) + \alpha^* \rho(F(\mathcal{X}), F(\mathcal{X}_i)) \qquad (1)$$

In the light of the previous discussion, we want α^* to be small (closer to 0.0) when novel and the corresponding gallery data have been acquired in similar illuminations, and large (closer to 1.0) when in very different ones. We show that α^* can be learnt as a function:

$$\alpha^* = \alpha^*(\mu), \qquad (2)$$

where μ is the *confusion margin* – the difference between the similarities of the two $\mathcal{X}_i$ most similar to $\mathcal{X}$. We compute an estimate of $\alpha^*(\mu)$ in a maximum a posteriori sense:

$$\alpha^*(\mu) = \arg\max_{\alpha} p(\alpha, \mu), \qquad (3)$$

where $p(\alpha, x)$ is the probability that α is the optimal value of the mixing coefficient.

2.2. Learning the α-Function

To learn the α-function defined in (3), we first need an estimate of the joint probability density $p(\alpha, \mu)$. The main

difficulty of this problem is of practical nature: in order to obtain an accurate estimate, a prohibitively large training database is needed. Instead, we propose a *heuristic* alternative, computed offline, from a small training corpus of individuals in different illumination conditions.

Our algorithm is based on an iterative incremental update of the density, initialized as a uniform density over the domain $\alpha, \mu \in [0, 1]$. We iteratively simulate matching of an unknown person against a set gallery individuals. In each iteration of the algorithm, these are randomly drawn from the offline training database. Since the ground truth identities of all persons in the offline database is known, we can compute the confusion margin $\mu(\alpha)$ for each $\alpha = k\Delta\alpha$, using the inter-personal similarity score defined in (1). Density $p(\alpha, \mu)$ is then incremented at each $(k\Delta\alpha, \mu(0))$ proportionally to $\mu(k\Delta\alpha)$.

The proposed offline learning algorithm is summarized in Fig. 2. A typical estimate of the probability density $p(\alpha, \mu)$ is shown in Fig. 3, with the corresponding α-function in Fig. 4.

3 Empirical Evaluation

Methods in this paper were evaluated on three databases:

- **FaceDB100**, with 100 individuals of varying age and ethnicity, and equally represented genders. For each person in the database we collected 7 video sequences of the person in arbitrary motion (significant translation, yaw and pitch, negligible roll), each in a different illumination setting, see Fig. 5 (a) and 6, at 10fps and 320×240 pixel resolution (face size ≈ 60 pixels).
- **FaceDB60**, kindly provided to us by Toshiba Corp. This database contains 60 individuals of varying age, mostly male Japanese, and 10 sequences per person. Each sequence corresponds to a different illumination setting, at 10fps and 320×240 pixel resolution (face size ≈ 60 pixels), see Fig. 5 (b).
- **FaceVideoDB**, freely available and described in [11]. Briefly, it contains 11 individuals and 2 sequences per person, little variation in illumination, but extreme and uncontrolled variations in pose and motion, acquired at 25fps and 160×120 pixel resolution (face size ≈ 45 pixels), see Fig. 5 (c).

Data acquisition: The discussion so far focused on recognition using fixed-scale face images. Our system uses a cascaded detector [16] for localization of faces in cluttered images, which are then rescaled to the unform resolution of 50×50 pixels (approximately the average size of detected faces in our data set).

Input: training data $D(person, illumination)$,
filtered data $F(person, illumination)$,
similarity function ρ,
filter F.
Output: estimate $\hat{p}(\alpha, \mu)$.

1: Init
$\hat{p}(\alpha, \mu) = 0$,

2: Iteration
for all illuminations $i,\ j$ **and** persons p

3: Initial separation
$\delta_0 = \min_{q \neq p} [\rho(D(p,i), D(q,j)) - \rho(D(p,i), D(p,j))]$

4: Iteration
for all $k = 0, \ldots, 1/\Delta\alpha,\ \alpha = k\Delta\alpha$

5: Separation given α
$\delta(k\Delta\alpha) = \min_{q \neq p} [\alpha\rho(F(p,i), F(q,j))$
$-\alpha\rho(F(p,i), F(p,j))$
$+(1-\alpha)\rho(D(p,i), D(q,j))$
$-(1-\alpha)\rho(D(p,i), D(p,j))]$

6: Update density estimate
$\hat{p}(k\Delta\alpha, \delta_0) = \hat{p}(k\Delta\alpha, \delta_0) + \delta(k\Delta\alpha)$

7: Smooth the output
$\hat{p}(\alpha, \mu) = \hat{p}(\alpha, \mu) * \mathbf{G}_{\sigma=0.05}$

8: Normalize to unit integral
$\hat{p}(\alpha, \mu) = \hat{p}(\alpha, \mu) / \int_\alpha \int_x \hat{p}(\alpha, x) dx d\alpha$

Figure 2. *Offline training algorithm.*

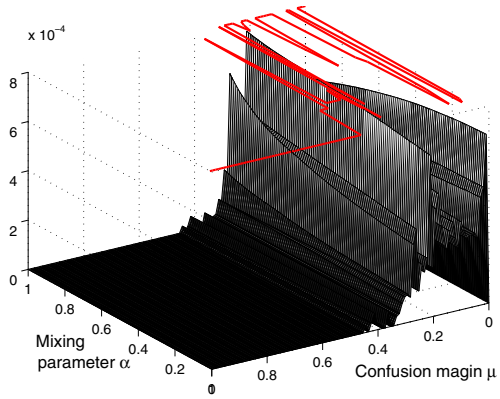
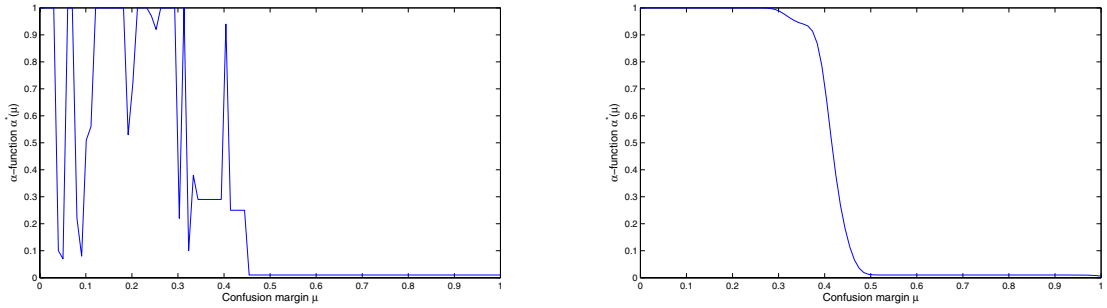


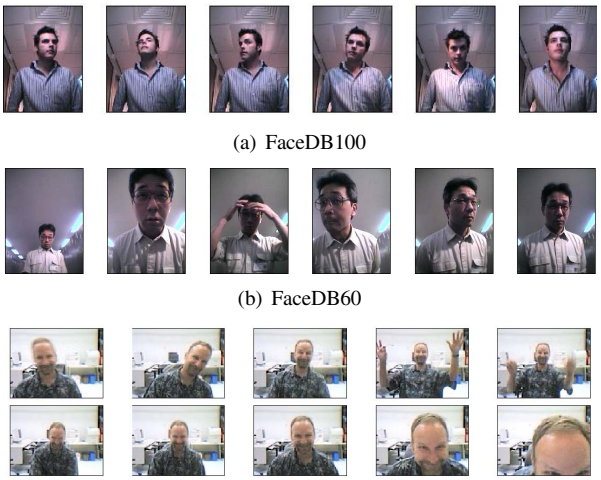
Figure 3. *Learnt probability density $p(\alpha, \mu)$ (greyscale surface) and a superimposed raw estimate of the α-function (solid red line) for a high-pass filter.*



(a) Raw $\alpha^*(\mu)$ estimate

(b) Smooth and monotonic $\alpha^*(\mu)$

Figure 4. *Typical estimates of the α-function plotted against confusion margin μ. The estimate shown was computed using 40 individuals in 5 illumination conditions for a Gaussian high-pass filter. As expected, α^* assumes low values for small confusion margins and high values for large confusion margins (see* (1)*).*

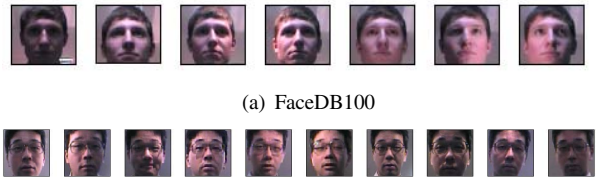


(a) FaceDB100

(b) FaceDB60

(c) FaceVideoDB

Figure 5. *Frames from typical video sequences from the 3 databases used for evaluation.*



(a) FaceDB100

(b) FaceDB60

Figure 6. *Different illumination conditions in databases FaceDB100 and FaceDB60.*

Figure 7. *Face representations evaluated.*

Methods and representations: The proposed framework was evaluated using the following filters (illustrated in Fig. 7):

- Gaussian high-pass filtered images [3, 8] (HP):

$$\mathbf{X}_H = \mathbf{X} - (\mathbf{X} * \mathbf{G}_{\sigma=1.5}), \tag{4}$$

- local intensity-normalized high-pass filtered images – similar to the Self-Quotient Image [17] (QI):

$$\mathbf{X}_Q = \mathbf{X}_H / (\mathbf{X} - \mathbf{X}_H), \tag{5}$$

  the division being element-wise,

- distance-transformed edge map [2, 6] (ED):

$$\mathbf{X}_E = \text{DistTrans}(\text{Canny}(\mathbf{X})), \tag{6}$$

- Laplacian-of-Gaussian [1] (LG):

$$\mathbf{X}_L = \mathbf{X} * \nabla\mathbf{G}_{\sigma=3}, \tag{7}$$

  and

- directional grey-scale derivatives [1, 7] (DX, DY):

$$\mathbf{X}_x = \mathbf{X} * \frac{\partial}{\partial x}\mathbf{G}_{\sigma_x=6}, \quad \mathbf{X}_y = \mathbf{X} * \frac{\partial}{\partial y}\mathbf{G}_{\sigma_y=6}. \tag{8}$$

As a baseline, to establish the difficulty of the evaluation data set, we compared the performance of our recognition algorithm to that of:

- State-of-the-art commercial system FaceIt® by Identix [12] (the best performing software in the most recent Face Recognition Vendor Test [13]),
- Constrained MSM (CMSM) [9] used in a state-of-the-art commercial system FacePass® [15],
- Mutual Subspace Method (MSM) [9], and

- KL divergence-based algorithm of Shakhnarovich *et al.* (KLD) [14].

In all tests, both training data for each person in the gallery, as well as test data, consisted of only a single sequence. Offline training of the proposed algorithm was performed using 40 individuals in 5 illuminations from the FaceDB100 – we emphasize that these were not used as test input for the evaluations reported in this section.

3.1. Results

We first evaluated the performance of the four established methods used for comparison purposes using raw greyscale input. A summary of the results is shown in Tab. 3.1. Firstly, note the poor performance of the KLD method. KLD can be considered as a proxy for gauging the difficulty of the recognition task, seeing that this algorithm can be expected to perform relatively well if the imaging conditions are not greatly different between training and test data sets [14]. This is further corroborated by observing that even the two best-performing methods, Identix's FaceIt and Toshiba's CMSM, incorrectly recognized about a quarter of individuals in our database. Interestingly, while performing marginally better, CMSM showed significantly less robustness to the particular data acquisition conditions used, as witnessed by its more than twice higher deviation of recognition scores across training/test combinations used.

Next, we evaluated the performance of CMSM and MSM using each of the 7 face image representations (raw input and 6 filter outputs). Recognition results for the 3 databases are shown in blue in Fig. 8 (the results on FaceVideoDB are tabulated in Fig. 8 (c), for the ease of visualization). Confirming the first premise of this work as well as previous research findings, all of the filters produced an improvement in average recognition rates. Little interaction between method/filter combinations was found, Laplacian-of-Gaussian and the horizontal intensity derivative producing the best results and bringing the best and average recognition errors down to 12% and 9% respectively.

In the last set of experiments, we employed each of the 6 filters in the proposed data-adaptive framework. Recognition results for the 3 databases are shown in red in Fig. 8. The proposed method produced a dramatic performance improvement in the case of all filters, reducing the average recognition error rate to only 4% in the case of CMSM/Laplacian-of-Gaussian combination.This is a very high recognition rate for such unconstrained conditions (see Fig. 5), small amount of training data per gallery individual and the degree of illumination, pose and motion pattern variation between different sequences. An improvement in the robustness to illumination changes can also be seen in the significantly reduced standard deviation of the recognition. Finally, it should be emphasized that the demonstrated improvement is obtained with a negligible increase in the computational cost as all time-demanding learning is performed offline.

Table 1. *Recognition rates (mean/STD, %).*

	FaceIt	CMSM	MSM	KLD
FaceDB100	64.1/9.2	73.6/22.5	58.3/24.3	17.0/ 8.8
FaceDB60	81.8/9.6	79.3/18.6	46.6/28.3	23.0/15.7
FaceVideoDB	91.9	91.9	81.8	59.1
Average	72.1	76.8	55.7	21.8

4. Conclusions

We described a novel framework for automatic face recognition in the presence of varying illumination, applicable to matching face sets or sequences, as well as to single shot-based recognition. Evaluated on a large, real-world data corpus, the proposed framework was shown to be successful in video-based recognition across a wide range of illumination, pose and face motion pattern changes.

References

[1] Y Adini, Y. Moses, and S. Ullman. Face recognition: The problem of compensating for changes in illumination direction. *PAMI*, 19(7), 1997.

[2] O. Arandjelović and R. Cipolla. Face recognition from video using the global shape-illumination manifold. *ECCV*, 2006.

[3] O. Arandjelović and A. Zisserman. Automatic face recognition for film character retrieval in feature-length films. *CVPR*, 2005.

[4] P. N. Belhumeur and D. J. Kriegman. What is the set of images of an object under all possible lighting conditions? *CVPR*, 1996.

[5] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. *SIGGRAPH*, 1999.

[6] J. Canny. A computational approach to edge detection. *PAMI*, 8(6), 1986.

[7] M. Everingham and A. Zisserman. Automated person identification in video. *CIVR*, 2004.

[8] A. Fitzgibbon and A. Zisserman. On affine invariant clustering and automatic cast listing in movies. *ECCV*, 2002.

[9] K. Fukui and O. Yamaguchi. Face recognition using multi-viewpoint patterns for robot vision. *International Symposium of Robotics Research*, 2003.

[10] A. S. Georghiades, D. J. Kriegman, and P. N Belhumeur. Illumination cones for recognition under variable lighting: Faces. *CVPR*, 1998.

[11] D. O. Gorodnichy. Associative neural networks as means for low-resolution video-based recognition. *International Joint Conference on Neural Networks*, 2005.

[12] Identix. Faceit. `http://www.FaceIt.com/`.

[13] P. J. Phillips, P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and J. M. Bone. FRVT 2002: Overview and summary. *Technical report, National Institute of Justice*, 2003.

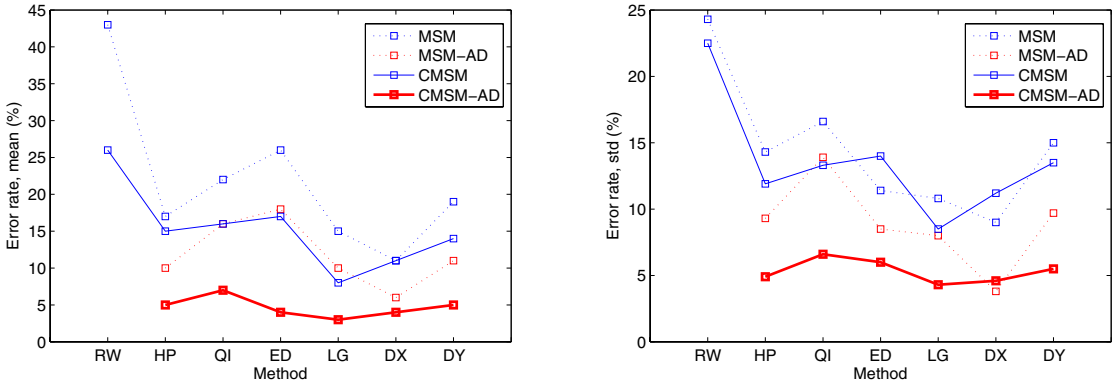
[14] G. Shakhnarovich, J. W. Fisher, and T. Darrel. Face recognition from long-term observations. *ECCV*, 3, 2002.

[15] Toshiba. Facepass. `www.toshiba.co.jp/mmlab/tech/w31e.htm`.

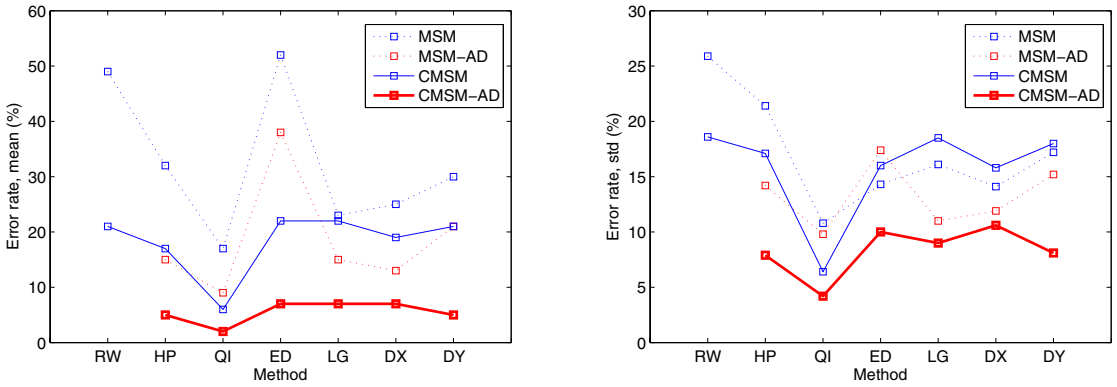
[16] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 57(2), 2004.

[17] H. Wang, S. Z. Li, and Y. Wang. Face recognition under varying lighting conditions using self quotient image. *AFG*, 2004.

[18] X. Wang and X. Tang. Unified subspace analysis for face recognition. *ICCV*, 2003.



(a) FaceDB100



(b) FaceDB60

	RW	HP	QI	ED	LG	DX	DY
MSM	0.00	0.00	0.00	0.00	9.09	0.00	0.00
MSM-AD	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CMSM	0.00	9.09	0.00	0.00	0.00	0.00	0.00
CMSM-AD	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(c) FaceVideoDB, mean error (%)

Figure 8. *Error rate statistics. The proposed framework (-AD suffix) dramatically improved recognition performance on all method/filter combinations, as witnessed by the reduction in both error rate averages and their standard deviations.*