2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance

Boosting Face Recognition in Real-World Surveillance Videos

Le An, Bir Bhanu, Songfan Yang
Center for Research in Intelligent Systems
University of California at Riverside, Riverside, CA 92507, USA
lan004@ucr.edu, bhanu@cris.ucr.edu, songfan.yang@email.ucr.edu

Abstract

Face recognition becomes a challenging problem in real-world surveillance videos where the low-resolution probe frames exhibit variations in pose, lighting condition, and facial expressions. This is in contrast with the gallery images which are generally frontal view faces acquired under controlled environments. A direct matching of probe images with gallery data often leads to poor recognition accuracy due to the significant discrepancy between the two kinds of data. In addition, the artifacts such as low resolution, blurriness and noise further enlarge this discrepancy. In this paper, we propose a video based face recognition framework using a novel image representation called warped average face (WAF). *The WAFs are generated in two stages: in-sequence warping and frontal view warping. The WAFs can be easily used with various feature descriptors or classifiers. As compared to the original probe data, the image quality of the WAFs is significantly better and the appearance difference between the WAFs and the gallery data is suppressed. Given a probe sequence, only a few WAFs need to be generated for the recognition purpose. We test the proposed method on the ChokePoint dataset and our in-house dataset of surveillance quality. Experiments show that with the new image representation, the recognition accuracy can be boosted significantly.*

1. Introduction

In recent years, video surveillance cameras have been widely deployed in public areas or residence premises for various purposes such as access control, security monitoring, etc. Although face recognition has been studied extensively, it is still very challenging for the existing face recognition algorithms to work accurately in reality [8].

Empirical studies have shown that the face image of approximate size 64×64 is required for existing algorithms to yield good results [15]. However, when a subject is not in the close vicinity of the camera, the captured face would have very low resolution. In addition, video sequences often suffer from motion blur and noise, together with changes in pose, lighting condition and facial expression.

Figure 1 shows some sample probe images from surveillance video data. Note that compared to the frontal faces in the gallery, the appearance of the probe images is quite different. How to tackle the discrepancy between the probe and gallery data becomes critical in developing a robust recognition algorithm.

Figure 1. Sample probe images (top) and gallery images (bottom) from the ChokePoint dataset [22].

To overcome the above limitations one strategy is to build a 3D face model to handle varying poses. Blanz and Vetter [6] generated a 3D morphable model as a linear combination of basis exemplars. The model was fit to an input image by changing the shape and albedo parameters of the model. Aggarwal and Harguess [9] used average-half-face instead of the whole face to improve the face recognition accuracy for 3D faces. Barreto and Li [12] proposed a framework for 3D face recognition system with variation of expression. The disadvantage for 3D based recognition is the high computational cost in building the 3D model. In addition, constructing a 3D model from low-resolution inputs is very difficult when the facial control points cannot be accurately localized by detectors.

To cope with the low-resolution issue in video based face recognition, Hennings-Yeomans *et al.* [10] used features from the face and super-resolution priors to extract a high-resolution template that simultaneously fit the super-resolution and face feature constraints. A generative model was developed in [3] for separating the illumination and down-sampling effects to match a face in a low-resolution video sequence against a set of high resolution gallery se-

978-0-7695-4797-8/12 $26.00 © 2012 IEEE
DOI 10.1109/AVSS.2012.17

quences. In [4] a person re-identification problem was addressed in distributed camera networks using facial appearance features. Stallkamp *et al.* [18] introduced a weighting scheme to evaluate individual contribution of each frame in a video sequence. In [17] face images with different modalities were projected into a common subspace for matching. Recently, Biswas *et al.* proposed a learning-based likelihood measurement to match high-resolution gallery images with probe images from surveillance videos [5]. A recent survey on face recognition in videos can be found in [16]. The performance of these methods generally degrades when applied to real-world surveillance data. In addition, the learning based methods may not be viable due to the insufficient training data that is available in reality.

In this paper, we propose a framework for video based face recognition by generating a new face image representations called warped average face (WAF) in two stages. In the *first stage* the warped face is generated using several consecutive face images from the original probe sequence. The original face images are averaged to obtain a first level representation. The original face images are then warped towards this representation. After warping, the warped faces are averaged again to generate a new representation at the next level. The process can be repeated and the evolving representation, as verified in the experiments, become more reliable with the improved face appearance. In the *second stage*, the generated face representation is further aligned towards the frontal view using a frontal face template. In such a way the appearance difference between the probe and the gallery data is reduced. From a large number of frames in the original sequence, only a few WAFs are necessary in the recognition step to achieve good recognition accuracy, which is favorable when computational cost is considered. We use SIFT flow [13] to warp the face images during the two stages described above. SIFT flow is a dense matching algorithm that uses SIFT features to find the pixel-to-pixel correspondences between two images at the scene level. Our work is inspired by the reference model based alignment for facial expression recognition [23] and the image averaging procedure for face recognition in photographs [11] which significantly improved the accuracy of a standard face-recognition algorithm. The proposed novel image representation, warped average face, can be used in any video based face recognition algorithms using different feature descriptors or classifiers.

The rest of the paper is organized as follows. Technical details are provided in Section 2. Section 3 is dedicated to the experimental results and Section 4 concludes this paper.

2. Technical Approach

Figure 2 gives an outline of the proposed method. The WAFs are generated from the face images in the original sequence in two stages: in-sequence face warping and frontal view face warping. SIFT flow is used in both stages.

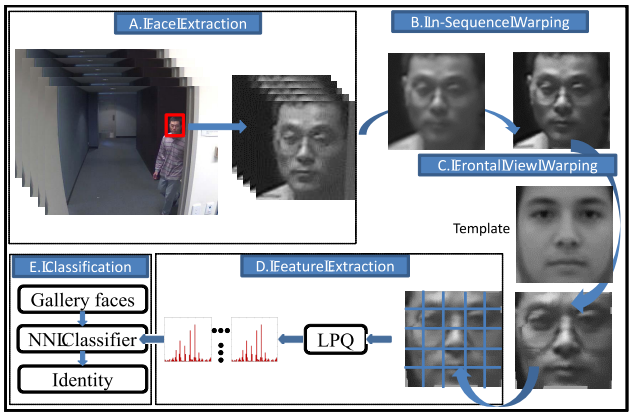

Figure 2. The overview of the proposed method.

2.1. SIFT Flow

SIFT flow was recently reported in [13] as an effective way to align images at the scene level. SIFT flow algorithm tries to match dense SIFT features [14] between two images and maintain spatial discontinuities. It is shown in [13] that scene pairs with high complexity can be robustly aligned.

First SIFT features for every pixel are extracted. Then similar to optical flow, an energy function is minimized to match two images s_1 and s_2:

$$E(\mathbf{w}) = \sum_{\mathbf{p}} min(\|s_1(\mathbf{p}) - s_2(\mathbf{p}+\mathbf{w}(\mathbf{p}))\|_1, t) + \quad (1)$$

$$\sum_{\mathbf{p}} \eta(|u(\mathbf{p})| + |v(\mathbf{p})|) + \quad (2)$$

$$\sum_{(\mathbf{p},\mathbf{q})\in\varepsilon} min(\alpha\,|u(\mathbf{p}) - u(\mathbf{q})|, d) + \\ min(\alpha\,|v(\mathbf{p}) - v(\mathbf{q})|, d) \quad (3)$$

where $\mathbf{p}$ denotes the elements on the image grid. $\mathbf{w}(\mathbf{p}) = (u(\mathbf{p}), v(\mathbf{p}))$ is the flow vector in horizontal and vertical direction. ϵ defines a local neighborhood. The term in (1) enforces the match along the flow vector $\mathbf{w}(\mathbf{p})$. (2) ensures the flow vector $\mathbf{w}(\mathbf{p})$ to be as small as possible without additional information. The smoothness constraint is imposed in (3) for the pixels in the local neighborhood. t and d are the thresholds for outliers and flow discontinuities. η and α are the scaling factors for the small displacement and smoothness constraint. To optimize this energy function, the dual-layer loopy belief propagation is used [13].

2.2. In-Sequence Face Warping

To reduce the variations among the consecutive frames in terms of pose, lighting, facial expression which adversely affect the recognition performance. We first perform in-sequence face warping in the first stage.

Given a face image $I(t)$ in a probe sequence, the initial averaged face (A_0) is computed by averaging all the face images in a local temporal window centered at t:

$$A_0 = \frac{1}{2k+1} \sum_{i=-k}^{k} I(t+i) \quad (4)$$

where k previous frames and k future frames together with the current frame $I(t)$ are averaged together.

In the next step, we use A_0 as a template to align the same $2k+1$ face images in the original sequence. The warped images are then averaged to obtain the warped face images at level A_1 by

$$A_1 = \frac{1}{2k+1} \sum_{i=-k}^{k} SIFT\_flow\left\langle I(t+i), A_0 \right\rangle \quad (5)$$

where $I(t+i)$ is warped towards A_0 by $SIFT\_flow$ function using A_0 as a template. In such a way the adjacent faces are warped towards a common appearance, the averaged output A_1 is sharper compared to A_0 due to the reduced frame variations. The quality of A_1 would degrade if the poses in the consecutive frames change significantly, in which case the SIFT flow will be more prone to error due to large out-of-plane rotations.

The generated A_1 can be further used as a matching template to generate A_2 at the next level. Higher levels can be achieved in the same manner. Some sample images of A_0, A_1, A_2 and A_3 are shown in Figure 3.

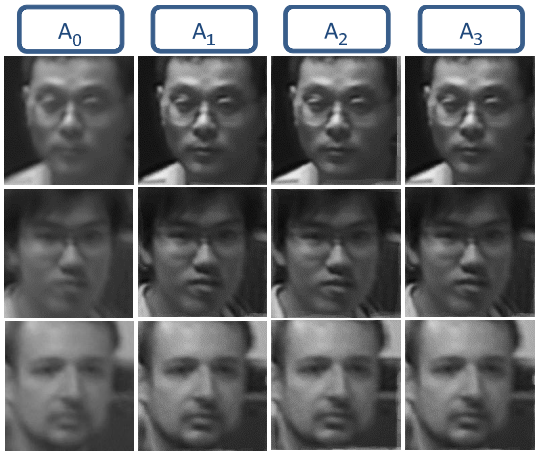

Figure 3. Examples of warped faces. From left to right: initial average face A_0, faces at level A_1, A_2, and A_3.

The face images at level A_1 contain more facial details compared to the initial averaged faces A_0. Moreover, the image noise are also suppressed in the averaging process due to the reason as follows. Assume the noise is Gaussian distributed $\mathcal{N}(\mu, \sigma^2)$, the averaging of $2k+1$ frames reduces the noise variance from σ^2 to $\sigma^2/(2k+1)$, given that the noise in different frames is independent and identically distributed. The blurriness is reduced by iteratively performing alignment and averaging.

From Figure 3 it is seen that the next two levels A_2 and A_3 are only slightly improved over A_1. The major reason is that the video sequences were captured at the frame rate of 30 fps, the variation between the adjacent frames is not significant. The variation is sufficiently minimized in A_1. Given a sequence of lower fps with larger motion between adjacent frames, we would expect the quality of the higher level images improves more aggressively.

We evaluate the image quality at different levels in term of blurriness using a blur metric [7] since blurriness is the major artifact at different representation levels. The blur metric is non-reference based. A scalar from 0 to 1 is calculated where 0 indicates a very sharp image.

Table 1 shows the average blurriness scores for a subset of the sequences of 25 subjects from the ChokePoint dataset [22]. The major score improvement takes place from A_0 to A_1 and the subsequent levels have minor improvement. The change in the quantitative scores corresponds to the visual examination in Figure 3. Taking into account the additional computational cost to obtain higher level representations, we choose to use A_1 in our experiments.

Table 1. Blurriness at different levels.

Level	A_0	A_1	A_2	A_3
Blurriness	0.7382	0.6451	0.5977	0.5924

The facial details are enhanced at this stage, which resembles the functionality of image super-resolution (SR). However, applying an image SR algorithm cannot achieve the same results neither in image quality nor in recognition performance (we compare our method with SR based method in the experiments).

2.3. Frontal View Face Warping

Although the generated A_1 image in the first stage is superior in image quality compared to the original probe data, directly matching the features extracted from A_1 faces and the gallery faces may lead to sub-optimal results due to the pose mismatch. SIFT flow is again applied here to warp the faces at level A_1 to a frontal view face template. The template is obtained by averaging the aligned frontal faces in the ChokePoint dataset and the FEI dataset [20] with 225 subjects in total. The reason for using an averaged template face is to avoid warping the faces towards the appearance of a specific person. With this process the WAF is generated with its appearance "forced" towards the frontal view.

Figure 4 shows examples of the generated WAFs. We use the face region instead of the whole head to eliminate the background influence. By aligning A_1 with the template, the output WAFs are "rotated" towards the frontal view. The pose variation between the WAF and the gallery image is decreased and the appearance of the two types of data becomes more coherent. Given a video sequence of N frames, approximately N/k WAFs are generated (k is about the half size of the temporal window as in Eq. 4). Note

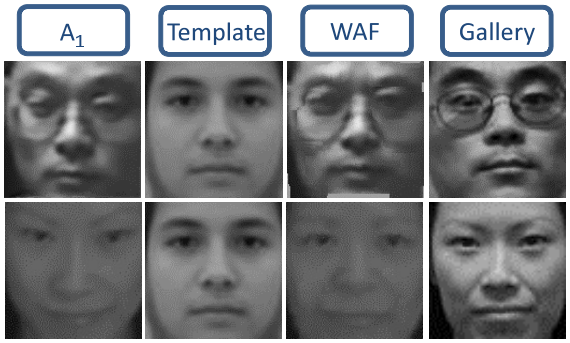

Figure 4. Examples of generated WAFs.

that different number of WAFs can be produced by defining how the time windows overlap when calculating the WAFs. The similarity between the appearances for the WAFs will increase if more common frames are used to compute the WAFs. Here we do not align the original probe images to the frontal view template directly in order to minimize the error in SIFT flow when matching low quality probe images to high quality gallery images.

The two-stage warping enhances the facial details, suppresses image blurriness and noise, reduces probe data variations, and corrects the pose of the probe data towards the desired frontal view in an integrated manner.

2.4. Feature Extraction and Classification

With the generated WAFs as the new face image representations, the features are then extracted from WAFs to describe the face. The commonly used face descriptors include local binary patterns (LBP) [1], local phase quantization (LPQ) [2], etc. In our application, the video captured by surveillance camera undergoes explicit blurriness (see Figure 1). LPQ has been applied to face recognition with success in handling blurred input [2]. Therefore, we choose LPQ as our face descriptor.

The basic idea of LPQ is that the phase of the original image and the blurred image has invariant property when the point spread function (PSF) is centrally symmetric. At each pixel location $\mathbf{x} = [x_1, x_2]^T$ in a neighborhood $N_\mathbf{x}$ of the image $f(\mathbf{x})$, the local spectra are computed using a discrete short time Fourier transform (STFT) by

$$F(\mathbf{u}, \mathbf{x}) = \sum_{\mathbf{y}} f(\mathbf{y}) w(\mathbf{y} - \mathbf{x}) e^{-j2\pi \mathbf{u}^T \mathbf{y}} \qquad (6)$$

where $w(\mathbf{x})$ defines the neighborhood $N_\mathbf{x}$. The local Fourier coefficients are computed at four frequency points and the phase information is recorded by a binary quantizer that observes the signs of the real and imaginary parts of each component in $\mathbf{F}(\mathbf{x})$. For each image block, a histogram is generated. The final descriptor of the whole image is obtained by concatenating all the histograms from different blocks. Note that in the proposed framework, any feature descriptors other than LPQ can be used.

After LPQ features are extracted from the WAFs, the Chi-square distance is used [1] to compute the feature distance. We apply a nearest-neighbor (NN) classifier. The distance scores are accumulated for all the WAFs generated from the original probe sequence and the lowest summed score provides the identity of the subject. In this case each WAF is considered equally important. A frame weighting scheme [18] can be applied to the WAFs to further improve the recognition performance.

3. Experiments

3.1. Datasets and Settings

We use ChokePoint dataset [22] which is designed for evaluating face recognition algorithms under real-world surveillance conditions. A subset of the video sequences (S1 and S2) from two cameras (C1 and C2) in portal 1 in two directions (E and L) are used (P1E_S1_C1, P1E_S1_C2, P1E_S2_C1, P1E_S2_C2, P1L_S1_C1, P1L_S1_C2, P1L_S2_C1, P1L_S2_C2). 25 subjects in total are involved. The captured faces in these sequences have large variations in pose, lighting condition, sharpness, background, etc (see Figure 1). The gallery set contains the high-resolution frontal faces of the 25 subjects. The extracted faces are provided with the dataset.

Figure 5. Sample probe images (top) and gallery images (bottom) from our Vface dataset.

In addition, our in-house dataset (Vface) involving 21 subjects is also collected for evaluation purpose. Frontal faces are stored as gallery set and we use two surveillance cameras (Vface_C1, Vface_C2) to record sequences for each subject walking naturally at a distance with uncontrolled lighting condition and background. The faces are detected using a trained Viola-Jones detector [21]. Figure 5 shows some sample probe and gallery data of our database. Note that as compared to the ChokePoint dataset, the probe images in our dataset have even lower quality in terms of resolution, pose variations and artifacts, which make our dataset more challenging. Both datasets well simulate the real-world case for surveillance purposes.

For both datasets, the probe faces are normalized to 64×64. For each sequence, the initial 20 frames are chosen to form a challenging problem since the subjects were far

Table 2. Recognition rate. The gains of the proposed method over baseline and baseline+SR methods are shown.

	P1E_S1_C1	P1E_S1_C2	P1E_S2_C1	P1E_S2_C2	P1L_S1_C1	P1L_S1_C2	P1L_S2_C1	P1L_S2_C2	Vface_C1	Vface_C2
Baseline	28%	44%	32%	48%	72%	56%	68%	56%	19.0%	14.3%
Proposed (gain)	+8%	+8%	+4%	+8%	+24%	+28%	+12%	+12%	+19.1%	+19.0%
Baseline+SR	28%	36%	40%	44%	52%	56%	76%	60%	9.5%	9.5%
Proposed (gain)	+8%	+16%	-4%	+12%	+44%	+28%	+4%	+8%	+28.6%	+23.8%

away from cameras in the beginning of the sequences. To generate WAF at the current frame, its previous and future 4 frames and itself are used (when the previous or future frame are not available, its mirror image with respect to the current frame is used, e.g., $I(t+1)$ is used when $I(t-1)$ is not available). In our method, we use 4 WAFs generated from the 20 frames at every fifth frame. The parameters for LPQ are set to $M = 7$,$\alpha = 1/7$ and $\rho = 0.9$. The size of the image block in LPQ computation is chosen as 16×16.

3.2. Experimental Results

We compare the results of our method with a baseline method where each original probe frame is used to match with the gallery images. The distance score for each frame is summed and the final identity is taken as the one with the lowest total score. We also compare with a SR based method (baseline + SR) by first super-resolving the probe image by a factor of 2 using a kernel regression based method [19]. We do not directly compare with the results on the ChokePoint dataset reported in [22] where a *video-to-video verification* protocol is used. The *video-to-image recognition* in our case is more challenging due to the significant data discrepancy between the probe and the gallery.

Table 2 shows the recognition rates for the probe sequences. Compared with the baseline and baseline + SR methods, the proposed method improves the recognition rate in all the testing sequences except P1E_S2_C1 where baseline + SR method is slightly better. The baseline + SR method is not helpful in improving the recognition accuracy. The reason is that in the the super-resolved images, the artifacts such as noise and blurriness are also magnified and the pose mismatch between the probe and the gallery data is accounted. With the proposed method, as the variations between the probe frames are alleviated and the appearance discrepancies between the probe and gallery images are reduced, the matching is less prone to error. This demonstrates the improvement using the WAFs over the baseline and baseline + SR method. The recognition rates are lower for Vface dateset. The reason is that in ChokePoint dataset the two cameras were mounted closely and the near-frontal views were captured, while in Vface dataset the cameras were more apart and the captured faces were less frontal.

Figure 6 shows some super-resolved images and comparison to A_1 images. As can be seen, the inherent noise and blurriness are not suppressed in the super-resolved images while A_1 contains more facial details and less noise.

The cumulative match curves (CMC) are given in Figure 7. For the same sequences in the same direction the results of two inputs by C1 and C2 are shown together. We are interested here to compare different methods for the same sequence. Note that for the same testing sequence, the recognition rates at different ranks are higher using the proposed method for most of the sequences. The baseline + SR method yields poor results for the Vface dataset. The main reason is that the SR reconstruction in the first step is not able to enhance the image quality due to severe blurriness and noise. While the face size is increased, the artifacts are also magnified. The accuracy gain by using WAF is larger for the Vface dataset, which indicates that the WAF is even more competitive in more challenging situations.

Figure 6. Samples of super-resolved faces (top) and A_1 (bottom).

As validated by the results, the novel face representations (WAF) help to boost the recognition accuracy while keeping the feature descriptor and the classifier unchanged.

4. Conclusions

To tackle the face recognition in surveillance video where the probe images exhibits variations in pose, lighting, expression as well as poor image quality, we propose a novel face image representation called warped average face (WAF) which contains enhanced facial details with fewer artifacts and corrected pose to be matched with the high-resolution frontal view gallery data. WAFs are generated in two stages by first warping the faces in the probe sequence and then aligning the warped face to a frontal view template. Only a few WAFs need to be extracted from the original sequence. Compared to the baseline method using the original probe images and the baseline method using super-resolved probe images, the proposed method achieves higher recognition rates both on a public dataset and our own dataset. The proposed method is simple yet effective and can be adapted to any video based face recognition schemes using different feature descriptors or classifiers. In the future, we will make our dataset publicly available and extend the proposed framework to multi-camera recognition scenarios.

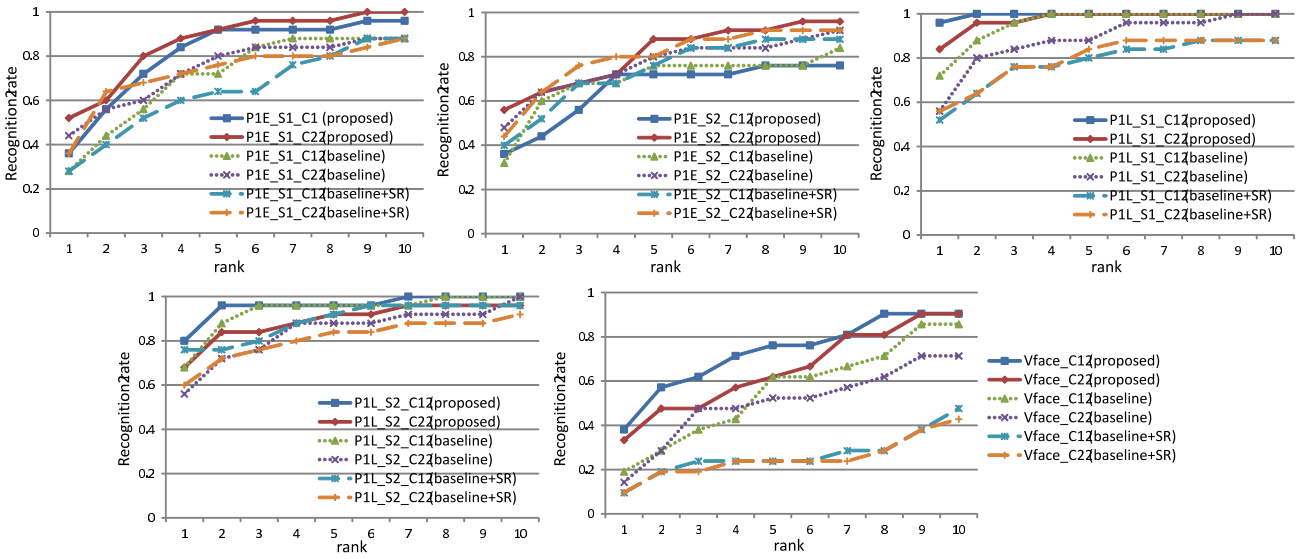

Figure 7. CMC for the testing sequences. The comparison is among different methods for the same sequence (e.g. P1E_S1_C1 proposed, P1E_S1_C1 baseline, P1E_S1_C1 baseline+SR). The inter-camera (C1 and C2) results are not to be compared directly.

Acknowledgment

This work was supported in part by NSF grant 0905671 and ONR grant on Aware Building.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face recognition with local binary patterns. In *Proc. ECCV*. 2004.

[2] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkila. Recognition of blurred faces using local phase quantization. In *Proc. ICPR*, 2008.

[3] O. Arandjelović and R. Cipolla. A manifold approach to face recognition from low quality video across illumination and pose using implicit super-resolution. *In Proc. ICCV*, 2007.

[4] M. Bauml, K. Bernardin, M. Fischer, H. Ekenel, and R. Stiefelhagen. Multi-pose face recognition for person retrieval in camera networks. In *Proc. AVSS*, 2010.

[5] S. Biswas, G. Aggarwal, and P. Flynn. Face recognition in low-resolution videos using learning-based likelihood measurement model. In *Proc. IJCB*, 2011.

[6] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE T-PAMI*, 2003.

[7] F. Crete, T. Dolmiere, P. Ladret, and M. Nicolas. The blur effect: perception and estimation with a new no-reference perceptual blur metric. *Proceedings of SPIE*, 2007.

[8] M. Grgic, K. Delac, and S. Grgic. Scface - surveillance cameras face database. *Multimedia Tools Appl.*, 2011.

[9] J. Harguess and J. Aggarwal. A case for the average-half-face in 2d and 3d for face recognition. In *Proc. CVPR Workshop*, 2009.

[10] P. Hennings-Yeomans, S. Baker, and B. Kumar. Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In *Proc. CVPR*, 2008.

[11] R. Jenkins and A. M. Burton. 100 % accuracy in automatic face recognition. *Science*, 319, 2008.

[12] C. Li and A. Barreto. An integrated 3d face-expression recognition approach. In *Proc. ICASSP*, 2006.

[13] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE T-PAMI*, 2011.

[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[15] Y. M. Lui, D. Bolme, B. Draper, J. Beveridge, G. Givens, and P. Phillips. A meta-analysis of face recognition covariates. In *Proc. BTAS*, 2009.

[16] C. Shan. Face recognition and retrieval in video. In D. Schonfeld, C. Shan, D. Tao, and L. Wang, editors, *Video Search and Mining*. Springer Berlin / Heidelberg, 2010.

[17] A. Sharma and D. Jacobs. Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. In *Proc. CVPR*, 2011.

[18] J. Stallkamp, H. Ekenel, and R. Stiefelhagen. Video-based face recognition on real-world data. In *Proc. ICCV*, 2007.

[19] H. Takeda, S. Farsiu, and P. Milanfar. Kernel regression for image processing and reconstruction. *IEEE T-IP*, 2007.

[20] C. E. Thomaz and G. A. Giraldi. A new ranking method for principal components analysis and its application to face image analysis. *Image Vision Comput.*, 2010.

[21] P. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 2004.

[22] 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 *Proc. CVPR Workshops*, 2011.

[23] S. Yang and B. Bhanu. Understanding discrete facial expressions in video using an emotion avatar image. *IEEE T-SMC B*, 2012.