Low-Resolution Video Face Recognition with Face Normalization and Feature Adaptation

Christian Herrmann [#*1], Chengchao Qu [#*2], Jürgen Beyerer [*#3]

[#] *Vision and Fusion Lab, Karlsruhe Institute of Technology KIT*
Adenauerring 4, Karlsruhe, Germany

[*] *Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB*
Fraunhoferstrasse 1, Karlsruhe, Germany
[1] christian.herrmann@iosb.fraunhofer.de
[2] chengchao.qu@iosb.fraunhofer.de
[3] juergen.beyerer@iosb.fraunhofer.de

Abstract—Face analysis is a challenging topic, especially when addressing low-resolution data. While face detection is working satisfactorily on such data, further facial analysis often struggles. We specifically address the issues of face registration, face normalization and facial feature extraction to perform low-resolution face recognition. For face registration, an approach for landmark detection, pose estimation and pose normalization is presented. In addition, a strategy to mirror the visible face half in the case of a rotated face is suggested. Next, the normalized face is used to extract the features for recognition. Using situation adapted local binary patterns (LBP) which are collected according to the proposed framework, including several scales and spatial overlaps, boosts the recognition performance well above the baseline. Results are presented on the YouTube Faces Database which is the current state-of-the-art dataset for video face recognition. Proper adjustments are made to convert this high-resolution dataset to a low-resolution one. We show that the presented adaptations increase face recognition performance significantly for low-resolution scenarios, closing a large part of the gap to high resolution face recognition.

I. INTRODUCTION

While high-resolution (HR) face analysis is well studied because of the availability of several large public datasets, low-resolution (LR) face analysis remains a challenge. However, the potential application domain is intriguing: The amount of surveillance footage is growing and exceeds more and more the manual analyzing capacities. This motivates the focus on video and LR data. Unlike TV program or movies where close-up shots are used to compensate the low resolution of video cameras, surveillance cameras usually are installed in a fixed position overlooking a larger area. Thus, faces with a larger distance to the camera lead to small face sizes in the image. The assumed scenario is motivated by forensic analysis of surveillance data, where large amounts of low quality material need to be inspected for the occurrences of a criminal. This might answer the question how a criminal arrived at a crime scene and left it afterwards, maybe identifying further accomplices. Searching for further occurrences based on one sample from the video footage requires a LR to LR face recognition system which we will describe here.

Basically, there are five processing steps in order to perform face recognition, also illustrated in figure 1: face detection/tracking, face registration, face normalization, face feature extraction and face feature comparison. While face detection and tracking is rather robust for low resolutions (approximately down to 20 pixels face size), the following steps are still struggling at such resolutions. Thus, we address these last steps in order to improve LR face recognition performance.

While HR face recognition is able to deliver impressive results, even surpassing the human performance on specific datasets [1], LR face recognition faces some additional challenges. According to [2] these are *misalignment*, *noise affection*, *lack of effective features* and *dimensional mismatch* between probe and gallery. In this contribution, we will specifically address two of these challenges, leading to an adaptation of the highlighted steps in figure 1. First, the face alignment by proposing a robust face normalization technique. Secondly, we adapt several features known from HR face recognition to the LR domain and show that they improve the effectiveness. With respect to noise affection, we assume that the given face size resembles the actual image quality and thus noise is negligible. In practical applications, the face resolution often exceeds the image quality, meaning that one can scale the face image down without loss of facial information. Thus, we assume that low quality face images are scaled down appropriately and have no significant quality flaws in the resulting resolution. Finally, dimensional mismatch is no major concern in our scenario which involves comparing LR faces to other LR faces.

The remaining part of this work is organized as follows. First, a robust strategy to register and normalize LR faces is presented in section III. This includes an innovative mirroring strategy in order to fill up invisible face parts in the case of large unconstrained head poses. Secondly, section IV explains how face recognition can be extended by specific features designed for low resolutions. Finally, large-scale experiments are performed in section V on the YouTube Faces Database (YTF).

II. RELATED WORK

Registration and normalization. Alignment of facial images is an indispensable preprocessing step for successful face recognition. Since the bounding boxes returned by modern face detection algorithms can tolerate up to a relatively large degree of variations in pose, expression
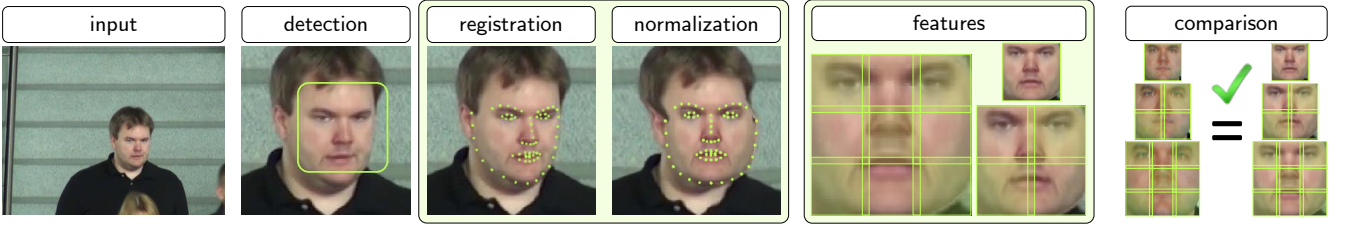
Fig. 1. Illustration of the steps from the source image to face comparison. The highlighted steps are addressed in this paper.

and occlusion, an extra registration or normalization can remove some of those degrees of freedom and leads to considerably better recognition performance. While image-based holistic registration techniques (e.g. Lucas-Kanade [3] or similarity transform [4]) suffice for some cases, model-based approaches have gained popularity in recent years including both 2D models [5], [6] and 3D models [7], [1]. Gao et al. [5] exploits the Active Appearance Model (AAM) [8] to map the facial texture back to the predefined shape-free coordinates by piecewise affine warping. Berg and Belhumeur [6] employ a Constrained Local Model (CLM) approach [9] to localize a few facial feature points and consequently conduct an identity-preserving warping based on the landmark locations of similar exemplars. With the help of a generic 3D face model, both Asthana et al. [7] and Taigman et al. [1] manage to normalize the frontal view by correspondence of the detected facial landmarks, providing aesthetically better results without the excessive use of morphable models [10].

In the context of LR face alignment, [11] and [12] explore different fitting strategies and features to alleviate the challenges caused by LR face images.

Recognition. Currently successful face recognition algorithms in the HR domain involve heavy machine learning. For example, by learning a Convolutional Neural Network [1] which implicitly trains the low level features and their combination to a discriminative face descriptor. Another example is [13], [14] where a large number of local features is collected across the face which are afterwards represented by a pre-trained cumulative descriptor.

LR face recognition has to work with limited data in two ways, thus learning based approaches are less popular. The first lack of data is problem inherent: the lower resolution means a lower number of pixels and thus less possibilities to collect features. The other data shortage comes from the lack of large scale datasets addressing this domain. This is mainly caused by the fact that labeling of LR face data is even hard for humans which leads to difficulties in generating the ground truth. Thus, LR face recognition work usually uses HR face datasets and scales them down [15], [16], [17], which is what we will perform too. A popular option to address the aspect of the low pixel numbers is to use no explicit features at all, meaning to work with the raw pixel values as model input. This has the advantage that manifold assumptions hold which offer an elegant way to model the face space [16], [17], [18]. When using explicit features, they are usually adapted to the LR domain, for example by temporal or scale based histogram fusion for the LBP descriptor [19]. A further way is to perform super-resolution

on the LR images and apply HR face recognition algorithms afterwards [20], [21].

III. FACE NORMALIZATION

The very first step of our normalization procedure is localizing the facial features on the image frame. By virtue of the recent advances of regression-based algorithms [22], [23], robust localization of landmarks, with relatively little influence from pose, expression, illumination, occlusion and resolution, can be achieved even under unconstrained conditions. 66 fiducial landmarks characterizing the main facial features are localized. Unlike Gao et al. [5] who warp the facial texture back to the shape-free frame, which loses the identity of shape after the normalization, we follow the same idea as Berg and Belhumeur [6] that preserves the individual characteristics of face shapes, but with a much simpler approach. The procedure of our face normalization approach is illustrated in figure 2.

In order to roughly estimate the facial pose, the original 2D landmarks are converted to 3D [24] by employing the 3D shape model trained by Saragih et al. [25]. The rotation matrix can then be obtained by least squares minimization of the projection of the corresponding 2D and 3D landmarks followed by imposing the orthonormality constraints on the projection matrix. Subsequently, the pitch, yaw and roll angles can be calculated.

Since the yaw rotation is one of the biggest nuisance factors in cross-pose face recognition, we take special care when normalizing pose, based on the yaw angle. If the angle is smaller than 10°, the image is regarded as a simple case where the problem of invisible texture as a result of self-occlusion is minor. In this case, we merely rotate the 3D shape back to frontal view and project the landmark locations onto 2D. After constructing the Delaunay triangulation of the landmarks, the texture values inside each triangle can be uniquely mapped to the target position by piecewise affine warping. As an example, the normalization result is depicted in the upper branch of figure 2. However, the visual quality of the normalized image is suboptimal due to a bit of large rotation angle.

As is analyzed in [26], when the head is performing yaw rotation, the detected 2D contour landmarks in the self-occluded face half continuously deviates from their actual position on the 3D model for lack of hidden texture representation in the sparse 3D model. This adds extra noise to the reconstructed 3D face model and causes inconsistency to the normalized frontal face. As is seen in figure 2, the eye and contour landmarks in the 'bad' half are stretched downward and outward, rendering a non-symmetric frontal
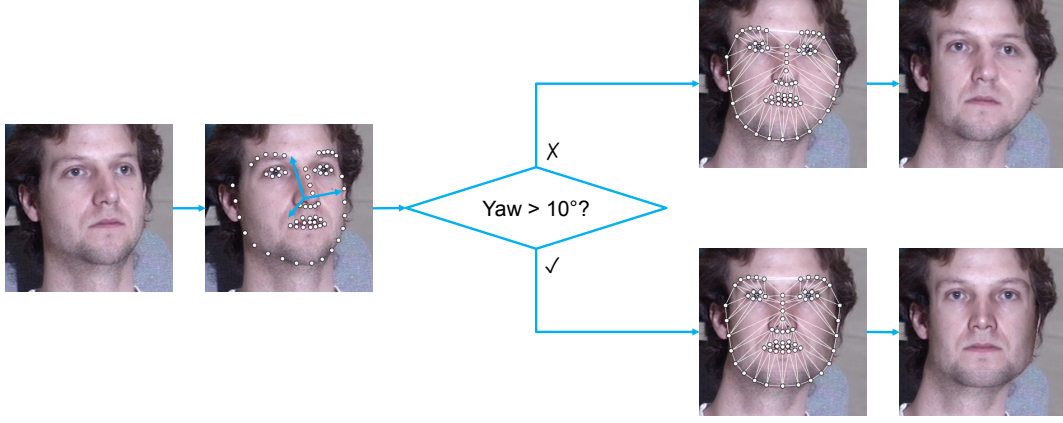
Fig. 2. Diagram of the proposed face normalization approach

face. Moreover, the nose and other landmarks in the middle line also tilt and these impacts become more severe with increasing yaw angle. To address this problem without losing the shape identity, we leverage the similarity property of human faces. Despite the low normalization quality, we notice that the landmarks in the 'good' face half are still reliable without the influence of self-occlusion. With this in mind, we first force the middle landmarks to form a perpendicular straight line based on their mean location. At the same time, the connected landmarks controlling the horizontal extent of nose and mouth are shifted respectively according to their relative position to the central line. Afterwards, the 'good' landmarks are mirrored to the 'bad' side to finalize the synthetic frontalization that accounts for facial symmetry. Last but not least, the texture quality of the 'bad' part is mostly very low, as it undergoes heavier warping and the precision of these landmarks is also insufficient. Hence, the texture is also copied along with the landmarks. The final result in the lower case of figure 2 demonstrates the efficacy of this approach, especially at the nose and stretched plain texture area.

IV. FACE RECOGNITION

When addressing low resolution face recognition, the most important step is to use an appropriate descriptor. Features used to build a HR descriptor require adjustment for LR scenarios [19]. For face comparison we will use two video face recognition systems. First, as representative of unsupervised approaches the Mutual Subspace Method (MSM) [27], [28] which models each face sequence by a linear subspace and compares them by the mutual principle angle. Secondly, we use the currently best-performing video face recognition approach that is trained on public data [13] which means it is a supervised approach. The key concept of that work is to collect dense SIFT-features across each frame, combine the features from all frames into one set and represent this feature set by the Fisher Vector. The final face track descriptor is constructed by a linear projection of the Fisher Vector to a low dimensional space. Comparison of the track descriptors is performed by the Euclidean distance. The main issue when using this strategy for LR face tracks is that it was designed for a face size of 150 pixels, including SIFT-features having a size of 24 pixels.

As the first step we perform an adaption of the feature extraction strategy by suggesting the framework illustrated in figure 3 that can be used to extract a set of local features from a LR face of size a. Extraction takes place on p scales of the scale-space pyramid and on each scale features are extracted in a regular grid-based manner. The relevant parameters are the region size r of each feature and the overlap v between regions, both measured in pixels.

Secondly, the scale-space image pyramid is adjusted to the LR scenario. Usually, the pyramid is consists of scales that differ by a constant scaling factor γ, thus the scaled face size is $a_{n+1} = \gamma \cdot a_n$. On LR images this leads to non-optimal coverage of the image area by the features. Because each feature is extracted from a fixed-size region of $r \times r$ pixels, scaling by a constant factor and feature extraction in a regular grid-based pattern leads to left over pixels at the image border. The impact of this discretization effect increases with a decreasing image size a. To prevent disregarding border pixels, we maximize the feature coverage on each scale by selecting the scaling factor appropriately. The result is illustrated in figure 3 on the right. In consequence, this leads to a linear scale-space pyramid instead of an exponential one where the scaled image size is $a_{n+1} = a_n + k$ instead of $a_{n+1} = \gamma \cdot a_n$. The linear adjustment k depends on the chosen feature size r and the overlap v as follows: $k = r - v$.

Thirdly, we propose to use LBP [29] instead of SIFT-features because their granularity better fits the LR scenario. The problem with SIFT-features is that they are computed on the derivative of the image, which is usually computed by applying a filter to the image (e.g. a Sobel filter). Thus, considering the filter size of at least 3×3 pixels, computing a $r \times r$ SIFT feature requires $(r + 1) \times (r + 1)$ underlying image pixels. While this can be neglected for high resolutions, it has a critical impact on the number of computable SIFT features for low resolution faces and small feature sizes r.

Finally, inspired by [13] we also include the option to collect flipped features, thus combining the features from the original image with the ones from the horizontally flipped image. Although no big benefit can be expected because of the previous application of the mirroring strategy for face normalization, the option is added for comparability.
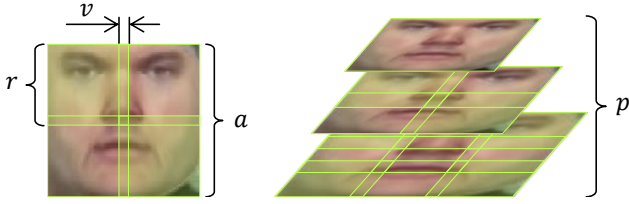
Fig. 3. Illustration of proposed feature extraction framework. From each region a local feature is extracted.

TABLE I
DECREASE IN DETECTION RATE FOR A VIOLA-JONES-BASED FACE
DETECTOR ON DOWNSCALED IMAGES. THE TRAINED MODEL HAS A
SIZE OF 20 PIXELS.

face size	14	20	26	32	38	HR
detection rate	0.027	0.897	0.970	0.979	0.987	1.000

V. EXPERIMENTS

The experiments are conducted on the YouTube Faces Database [30] which is a public in-the-wild dataset consisting of face videos collected on YouTube. This HR dataset is chosen because of the lack of according public LR datasets. LR faces are generated by downscaling to the resolution stated at each experiment. Motivated by the scenario of searching a specific person in the video data, evaluation uses an retrieval evaluation strategy with the mean average precision map as measure for face recognition performance [19], [31]. This measure has its origins in the area of information retrieval and rates the quality of a search result by the ranks of the correct samples. If $map = 1$ all correct samples are located at the top of the ranking.

YTF consists of 3,425 face videos with a typical face size larger than 100 pixels in width. Thus, the evaluation is performed on downscaled data starting with a resolution of 14×14 pixels upto 38×38 pixels. The lower bound appears to be the range where face analysis performance degrades heavily, whereas the upper bound is in the range where LR scenarios no longer apply. The dataset includes two sets of images, the original unregistered images including only the face detection bounding boxes, and a set of face images that are registered, normalized and cropped. The data of the first set is used in all the experiments whereas the second one is used as 'HR YTF normalized' data to compare the LR results to a HR solution.

A. Detection

In the introduction, we claimed that face detection is able to work at low resolutions. To prove this point, the percentage of successful face detections on the dataset is evaluated in dependence of the resolution. Note that the HR faces (> 100 pixels in this case) have already been detected by a Viola-Jones-based face detector [32]. Thus, a similar detector is used for low resolution experiments. Although some performance degradation can be observed in table I, it only drops significantly for detection sizes smaller than the trained model, which is obvious.

B. Normalization

The benefit of face registration and normalization is evaluated by its effect on the face recognition results. Similar
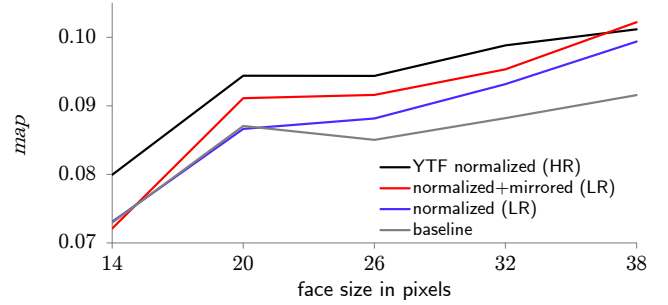

Fig. 4. Comparison of face recognition results when using the unprocessed detections (baseline), the proposed LR normalization with and without mirroring, and the HR YTF normalization

to [29], $LBP_{8,1}^{u2}$ features are extracted in regions of $r \times r$ pixels and concatenated to the final image face descriptor. The regions size r is optimized for each resolution. Face tracks are described by the MSM [27], [28] as baseline face recognition method. To keep the results clear, the further face recognition approach will be evaluated in the next section.

The proposed LR face normalization strategy is compared to the baseline without normalization and the HR YTF face normalization in figure 4. HR YTF normalization means to use the normalized HR faces given by the authors and scale them down to the according resolution afterwards. This is to indicate the expected upper bound for the proposed approach. Obviously, the HR results beat the baseline ones on all resolutions. The proposed LR normalization without the mirroring strategy improves the results above a face size of 26 pixels. On smaller face sizes landmark detection fails sometimes, therefore positive and negative effects equalize each other. When adding the mirroring strategy of the visible face half, baseline results are already beaten significantly at 20 pixels face size. Especially, the mirroring step further improves the recognition results compared to the LR normalization. Figure 5 illustrates the qualitative benefit of the mirroring strategy.

Performing the experiments shown by figure 4 with the official verification protocol proposed by the authors of YTF [30] results in accuracies between 0.607 and 0.639 with standard deviations in the order of 0.02. The large, and thus insignificant, standard deviation is the reason we decide for the more complex retrieval protocol and the map measure: with the help of the larger statistical base, significant differences between the approaches can be proven in most cases by a randomization test [33] with $\alpha = 0.05$. Table II holds the results of the test and shows where measured differences are significant. Summarized, this means that the proposed strategy is significantly better than the baseline for all face sizes of 20 pixels and above.

C. Recognition

Supported by the results of the last section, the LR normalization and mirroring strategy is to be used in all following experiments. First, the framework parameters including the region size r, the overlap v and the number of pyramid levels p are optimized for each resolution. Table III shows this process exemplary for the fisher vector based

Fig. 5. Qualitative results of the proposed normalization and mirroring strategy. The top row shows the input images and the detected landmarks, the center row shows the images after normalization and the bottom row after normalization and mirroring. Columns depict a face size of 14, 20, 26 and 32 pixels from left to right.

TABLE II

SIGNIFICANCE OF RESULTS ILLUSTRATED IN FIGURE 4. ONLY SIGNIFICANTLY WORSE MEASUREMENTS ARE DENOTED, FOR EXAMPLE THE LETTERS L,B IN THE ROW OF NORMALIZED (HR) AND FOR FACE SIZE 26 MEAN THAT THE HR NORMALIZATION RESULTS FOR THIS RESOLUTION ARE SIGNIFICANTLY BETTER THAN THE BASELINE AND LR NORMALIZATION ONES, BUT THE DIFFERENCE TO THE MIRRORING STRATEGY IS INSIGNIFICANT BECAUSE NO M IS LISTED.

name	abbr.	face size in pixels				
		14	20	26	32	38
YTF normalized (HR)	H	M,L,B	L,B	L,B	L,B	B
norm.+mirrored (LR)	M		B	B	B	B
normalized (LR)	L				B	B
baseline	B					

TABLE III

OPTIMIZATION OF THE FEATURE COLLECTION FRAMEWORK. CASE NUMBERS IN THE 'SIG.' COLUMN DENOTE CASES IN THE SAME BUNCH OF SETTINGS THAT RESULTED IN SIGNIFICANTLY WORSE RESULTS ACCORDING TO A RANDOMIZATION TEST.

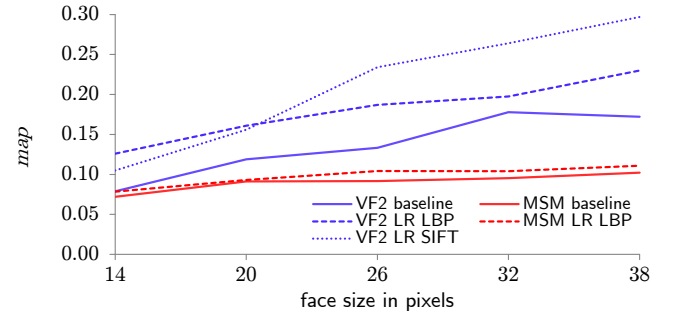| case	r	v	p	settings	LBP $a = 26$	sig.	SIFT $a = 26$	sig.
1				HR [13]	-		0.128	
2	6				0.132	4,5	-	
3	8				**0.136**	4,5	**0.199**	5
4	10				0.123	5	-	
5	12				0.115		0.183	
6	8	0			0.136		0.199	
7	8	2			0.169	6	0.212	6
8	8	4			**0.174**	6	**0.234**	6,7
9	8	2	3, exp		0.173		0.216	
10	8	2	5, exp		0.171		0.212	
11	8	2	3, lin		0.184	9,10	0.220	10
12	8	2	5, lin		**0.187**	9,10	**0.224**	9,10
13	8	2	5, lin	flip	0.189		0.221	

Fig. 6. Comparison of face recognition results when using the baseline feature extraction method used in section V-B and the proposed LR framework with either LBP or SIFT features.

VF2-algorithm [13] and a face size of $a = 26$ pixels. Specifically, it is shown that the feature extraction framework improves the results (cases 2 to 8) compared to the baseline and the linear pyramid excels the exponential one (cases 9 to 12). Increasing the overlap beyond case 8 is prohibited by the memory capacity of the test system. Accordingly, the pyramid was built upon case 7 as this again increases the number of features and combining the pyramid with the high overlap of case 8 also exceeds the memory capacity. Adding flipped features as suggested by [13] (case 13), has no significant effect on the results. This is because flipped features are only beneficial in cases where the head is turned which we already covered by the face mirroring. Note that LBP are worse than SIFT features at this scale, however, this is vice versa for smaller scales where discretization effects get worse as is shown by figure 6. Using the optimal value for each resolution leads to the results shown in figure 6. For both face comparison algorithms, MSM and VF2, the optimized features improve the results significantly. To relate the results again to the more popular verification protocol, we produce verification results with the same settings of the maximum value $map = 0.297$ which results in an accuracy of 0.816 ± 0.016. In comparison, the maximum verification accuracy achieved by [13] on HR data is 0.847 ± 0.014. Especially, we want to note that the achieved LR performance for a face size of 32 pixels and above exceeds the state-of-

the-art HR results from only two years ago on this dataset (accuracy of 0.789 ± 0.019 for a face size of 100 pixels [34]).

VI. CONCLUSION

Addressing the specific challenges of face recognition in LR video data showed that there is a potential to improve LR face recognition by adapting HR face recognition strategies. The most promising steps of face normalization and feature extraction were addressed and the results showed significant improvements by the proposed approaches. While very low-resolution face analysis addressing face sizes in the order of 15 pixels remains a significant challenge, it was shown that for face sizes above 20 pixels competitive results are possible. The achieved LR performance lies in the range of HR approaches which were presented only two years ago.

REFERENCES

[1] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[2] Z. Wang, Z. Miao, Q. M. J. Wu, Y. Wan, and Z. Tang, "Low-resolution face recognition: a review," *The Visual Computer*, vol. 30, no. 4, pp. 359–386, 2014.

[3] B. D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," in *Artificial Intelligence*, 1981, pp. 674–679.

[4] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, "Toward a Practical Face Recognition System: Robust Alignment and Illumination by Sparse Representation," *Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 372–386, Feb 2012.

[5] H. Gao, H. K. Ekenel, and R. Stiefelhagen, "Pose Normalization for Local Appearance-Based Face Recognition," in *International Conference on Biometrics*, 2009, pp. 32–41.

[6] T. Berg and P. N. Belhumeur, "Tom-vs-Pete Classifiers and Identity-Preserving Alignment for Face Verification." in *British Machine Vision Conference*, 2012.

[7] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. Rohith, "Fully automatic pose-invariant face recognition via 3D pose normalization," in *International Conference on Computer Vision*, 2011, pp. 937–944.

[8] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *European Conference on Computer Vision*, vol. 1407, 1998, pp. 484–498.

[9] P. N. Belhumeur, D. W. Jacobs, D. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *Computer Vision and Pattern Recognition*, 2011, pp. 545–552.

[10] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1063–1074, 2003.

[11] G. Dedeoğlu, S. Baker, and T. Kanade, "Resolution-Aware Fitting of Active Appearance Models to Low-Resolution Images," in *European Conference on Computer Vision*, vol. 2, 2006, pp. 83–97.

[12] C. Qu, E. Monari, and T. Schuchert, "Resolution-aware Constrained Local Model with mixture of local experts," in *Advanced Video and Signal Based Surveillance Workshops*, 2013, pp. 454–459.

[13] O. M. Parkhi, K. Simonyan, A. Vedaldi, and A. Zisserman, "A Compact and Discriminative Face Track Descriptor," in *Computer Vision and Pattern Recognition*, 2014.

[14] C. Herrmann and J. Beyerer, "Face Retrieval on Large-Scale Video Data," in *Computer and Robot Vision*, 2015, pp. 192–199.

[15] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *Computer Vision and Pattern Recognition*, 2010.

[16] K. Lee, J. Ho, M. Yang, and D. Kriegman, "Video-Based Face Recognition Using Probabilistic Appearance Manifolds," *Computer Vision and Pattern Recognition*, vol. 1, pp. 313–320, 2003.

[17] ——, "Visual Tracking and Recognition Using Probabilistic Appearance Manifolds," *Computer Vision and Image Understanding*, vol. 99, no. 3, pp. 303–331, 2005.

[18] R. Wang, S. Shan, X. Chen, and W. Gao, "Manifold-manifold distance with application to face recognition based on image set," in *Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[19] C. Herrmann, "Extending a local matching face recognition approach to low-resolution video," in *Advanced Video and Signal Based Surveillance*, 2013.

[20] W. W. Zou and P. C. Yuen, "Very low resolution face recognition problem," *Transactions on Image Processing*, vol. 21, no. 1, pp. 327–340, 2012.

[21] C. Qu, C. Herrmann, E. Monari, T. Schuchert, and J. Beyerer, "3D vs. 2D: On the Importance of Registration for Hallucinating Faces under Unconstrained Poses," in *Computer and Robot Vision*, 2015.

[22] X. Xiong and F. De la Torre, "Supervised Descent Method and Its Applications to Face Alignment," in *Computer Vision and Pattern Recognition*, 2013, pp. 532–539.

[23] C. Qu, H. Gao, E. Monari, J. Beyerer, and J.-P. Thiran, "Towards Robust Cascaded Regression for Face Alignment in the Wild," in *Computer Vision and Pattern Recognition Workshops*, 2015.

[24] I. Matthews, J. Xiao, and S. Baker, "2D vs. 3D Deformable Face Models: Representational Power, Construction, and Real-Time Fitting," *International Journal of Computer Vision*, vol. 75, no. 1, pp. 93–113, 2007.

[25] J. M. Saragih, S. Lucey, and J. F. Cohn, "Face alignment through subspace constrained mean-shifts," in *International Conference on Computer Vision*, 2009, pp. 1034–1041.

[26] C. Qu, E. Monari, T. Schuchert, and J. Beyerer, "Fast, robust and automatic 3D face model reconstruction from videos," in *Advanced Video and Signal Based Surveillance*, 2014, pp. 113–118.

[27] O. Yamaguchi, K. Fukui, and K. Maeda, "Face Recognition Using Temporal Image Sequence," in *Automatic Face and Gesture Recognition*, 1998.

[28] K. Fukui and O. Yamaguchi, "Face Recognition Using Multi-viewpoint Patterns for Robot Vision," *Robotics Research*, pp. 192–201, 2005.

[29] T. Ahonen, A. Hadid, and M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition," *Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[30] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Computer Vision and Pattern Recognition*, 2011.

[31] C. Herrmann and J. Beyerer, "Pyramid Mean Representation of Image Sequences for Fast Face Retrieval in Unconstrained Video Data," in *Lecture Notes in Computer Science 8888, Advances in Visual Computing, 10th International Symposium, ISVC 2014, Proceedings, Part II*, 2014, pp. 304–314.

[32] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[33] M. D. Smucker, J. Allan, and B. Carterette, "A comparison of statistical significance tests for information retrieval evaluation," in *Information and Knowledge Management*, 2007.

[34] L. Wolf and N. Levy, "The svm-minus similarity score for video face recognition," in *Computer Vision and Pattern Recognition*, 2013.