Robust Face Hallucination Using Quantization-Adaptive Dictionaries

Reuben Farrugia, Christine Guillemot

ROBUST FACE HALLUCINATION USING QUANTIZATION-ADAPTIVE DICTIONARIES

Reuben A. Farrugia *

University of Malta
Msida, Malta

Christine Guillemot

INRIA
Rennes-Bretagne-Atlantique, France

ABSTRACT

Existing face hallucination methods are optimized to super-resolve uncompressed images and are not able to handle the distortions caused by compression. This work presents a new dictionary construction method which jointly models both distortions caused by down-sampling and compression. The resulting dictionaries are then used to make three face super-resolution methods more robust to compression. Experimental results show that the proposed dictionary construction method generates dictionaries which are more representative of the low-quality face image being restored and makes the extended face hallucination methods more robust to compression. These experiments demonstrate that the proposed robust face hallucination methods can achieve Peak Signal-to-Noise Ratio (PSNR) gains between 2 – 4.48dB and recognition improvement between 2.9 – 8.1% compared with the low-quality image and outperforming traditional super-resolution methods in most cases.

Index Terms— Denoising, dictionary construction, face hallucination, face restoration, super-resolution.

1. INTRODUCTION

Closed Circuit Television (CCTV) systems are ubiquitous in many cities around the globe. These cameras are normally installed to cover a large field of view where the query face image may not be sampled densely enough by the camera sensors. Moreover, the captured footage is degraded by lossy image or video compression which reduces the texture details and generates block or ringing artefacts. The low-resolution and poor quality of the footage reduces the effectiveness of CCTV to identify perpetrators and potential eye witnesses [1].

Several learning-based face super-resolution methods [2–10] were proposed to increase the resolution of low-resolution face images. These methods use coupled low- and high-resolution dictionaries to learn mapping relations to hallucinate a high-resolution face from the observed low-resolution image [11]. However, the employed dictionaries do not consider the distortions caused by compression and are therefore brittle to artefacts which are not properly modelled.

Reconstruction-based super-resolution methods tried to include the quantization error caused by image/video compression in the image acquisition model [12, 13]. However, reconstruction based methods fail to achieve high-quality images at larger magnification factors. Moreover, these methods do not exploit the prior knowledge of the facial structure in the restoration process. Alternatively, compression artefacts can be suppressed using codec-dependent de-blocking algorithms [14–17]. Nevertheless, these methods still do not exploit the facial structure as prior, resulting in sub-optimal restored images which lack texture details crucial for recognition.

In this paper we introduce a more realistic image acquisition model for CCTV images which models the joint contribution of down-sampling and the distortions caused by compression. This model leads to the design of a dictionary contraction strategy which exploits the quantization parameter contained within the image file to be restored to derive low-quality dictionaries which are more representative of the image being restored. Three face hallucination methods, Position Patches (PP) [6], Sparse Position Patches (SPP) [7] and Multilayer Locality Constrained Iterative Neighbour Embedding (M-LINE) [10] were extended to make them more robust to compression artefacts. Experimental results show that the proposed Robust PP (RPP) accomplishes the best reconstruction ability achieving PSNR gains up to 4.45dB compared to PP. Moreover, the proposed Robust SPP (RSPP) and Robust M-LINE (RM-LINE) achieved comparable performance (most of the time superior) to SPP and M-LINE, achieving recognition improvements of up to 16.2%.

The remainder of the paper is organized as follows. The image acquisition model is introduced in Sec. 2 followed by the proposed dictionary construction method. In Sec. 4 we extend three face hallucination methods to make them robust to compression artefacts followed by a complexity analysis of the proposed methods. The experimental set-up was presented in Sec. 6, followed by the experimental results in Sec. 7 and concluded in Sec. 8 with the final remarks.

2. IMAGE ACQUISITION MODEL

Let $\mathbf{X}$ and $\mathbf{Y}$ denote the high- and low- resolution facial images, respectively. Many super-resolution methods formulate

*The first author performed the work while at INRIA Rennes-Bretagne-Atlantique, France.

the acquisition model as

$$\mathbf{X} = \downarrow_{\alpha} \mathbf{BY} + \boldsymbol{\eta} \tag{1}$$

where $\downarrow_{\alpha}$ represents a downscaling by a factor $\frac{1}{\alpha}$, $\mathbf{B}$ is a blurring function and $\boldsymbol{\eta}$ represents additive noise. However, this formulation is valid for uncompressed images and is not effective when $\mathbf{X}$ is compressed. The compression process introduces quantization error, which is the dominant source of errors when using large compression ratios which are typically used in CCTV systems.

In this work we use an acquisition model similar to the one proposed in [12, 13] for reconstruction based super-resolution where we ignore the motion-compensation component since single-image face hallucination is of interest here. The acquisition model considered in this work is given by

$$\widehat{\mathbf{X}} = \mathbf{T}^{-1} Q\left(\mathbf{T}\left[\downarrow_{\alpha} \mathbf{BY} + \boldsymbol{\eta}\right]\right) \tag{2}$$

where $\widehat{\mathbf{X}}$ is the decoded low-quality image, $Q()$ represents the non-linear quantization process (forward and inverse quantization) and $\mathbf{T}$ and $\mathbf{T}^{-1}$ are the forward and inverse-transform operations respectively.

3. DICTIONARY CONSTRUCTION

The resolution of the low-quality face image $\widehat{\mathbf{X}}$, which is blurred, downscaled and compressed, is defined by the distance between the eye centres d_x. The aim of the proposed method is to up-scale the face image $\widehat{\mathbf{X}}$ by a scale factor $\alpha = \frac{d_y}{d_x}$, where d_y represents the distance between the eye centres of the desired restored face image. The low-quality face image $\widehat{\mathbf{X}}$ is divided into overlapping patches of size $\sqrt{n} \times \sqrt{n}$ with an overlap of γ_x, and the resulting patches are reshaped to column-vectors $\widehat{\mathbf{x}}_i$, where $i \in [1, p]$ is the patch index.

A set $\mathcal{H}$ of m high-quality face images, which are registered using affine transformation computed on landmark points of the eyes and mouth center coordinates, were used to construct the high-quality dictionary, where the distance between the eye centres is set to d_y. These face images are divided into overlapping patches of size $[\alpha\sqrt{n} \times \alpha\sqrt{n}]$ with an overlap of $\gamma_y = [\alpha\gamma_x]$, where $[*]$ stands for the rounding operator. The i-th patch of every high-quality image is reshaped to a column-vector and placed within the high-quality dictionary of the i-th patch $\mathbf{H}_i$.

Existing learning-based face hallucination methods generate the low-quality dictionary using the strategy depicted in Fig. 1a, where each high-quality face image is blurred and downscaled by a factor $\frac{1}{\alpha}$ to generate a set of m low-quality face images $\mathcal{L}$. Every image in $\mathcal{L}$ is divided into overlapping patches of size $\sqrt{n} \times \sqrt{n}$ with an overlap of γ_x and each patch is vectorized and placed within the low-quality dictionary $\mathbf{L}_i$.

These methods however are based on the acquisition model defined in (1) which does not cater for the distortion caused by compression. On the other hand, the proposed method uses the acquisition model defined in (2) and extends the existing dictionary construction method by including the codec used to compress $\mathbf{X}$ within the degradation model[1] (see Fig. 1b). This approach is based on the assumption that since facial images have similar structure, registered facial images with the same resolution d_x will contain similar quantization noise patterns when compressed using the same codec configured with the same syntax element (SE)[2]. It then encodes all the down-sampled face images contained in $\mathcal{L}$ using the same configuration (based on the information contained within SE) used to encode $\mathbf{X}$ to derive the set of distorted face images $\widehat{\mathcal{L}}$. Every image contained within $\widehat{\mathcal{L}}$ is divided into patches of size $\sqrt{n} \times \sqrt{n}$ with an overlap of γ_x, where the resulting i-th patch is vectorized and placed within the refined low-quality dictionary $\widehat{\mathbf{L}}_i$. We emphasize here that in this work $\widehat{\mathbf{L}}_i$ is adapted based on the quantization parameter, but can be extended to exploit other syntax information.

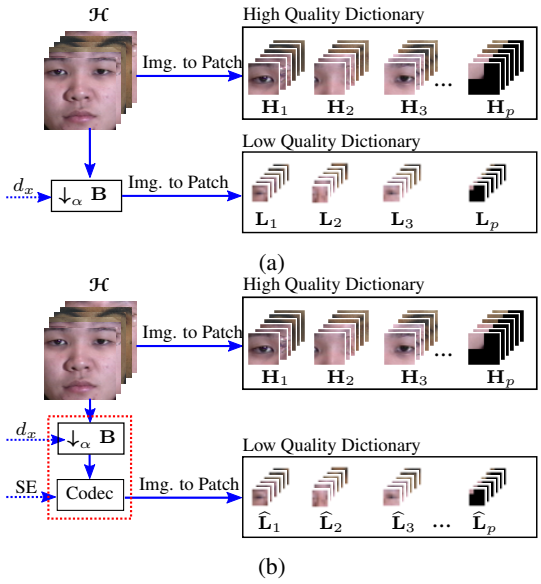

Fig. 1: Illustration of the a) traditional dictionary construction and b) proposed dictionary construction methods.

4. ROBUST FACE HALLUCINATION

Fig. 2 illustrates the schematic diagram of the proposed robust face hallucination method, where the low-quality face image $\widehat{\mathbf{X}}$ is divided into overlapping patches as described in Sec. 3. A learning based face hallucination method is employed to super-resolve the low-quality patches $\widehat{\mathbf{x}}_i$ to restore $\widetilde{\mathbf{y}}_i$, which are then combined by averaging overlapping pixels. The major contribution resides in the refined dictionary $\widehat{\mathbf{L}}_i$ (see Sec. 3) which uses the syntax information contained within the file to be restored, to derive a more representative dictionary. The proposed method is agnostic of the

[1]The degradation model is bounded by a dotted red box within Fig. 2.

[2]The syntax element contains the encoder's configuration parameters used to compress the source image.

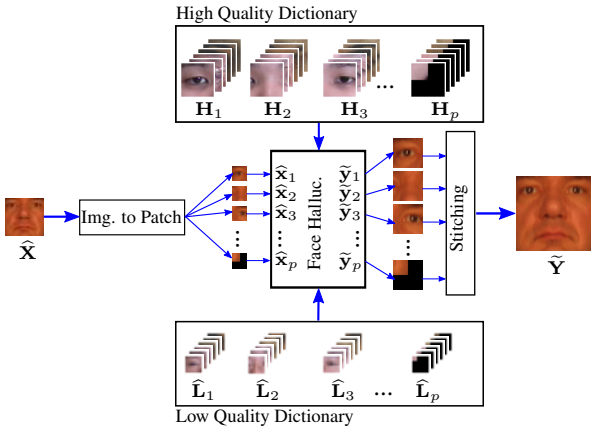

Fig. 2: Schematic diagram of the proposed Robust Face Hallucination method.

learning-based face hallucination and codec used. This work extends the formulation of three face hallucination methods which represent the state-of-the-art in super-resolving uncompressed facial images, where these extensions are summarized in the following sub-sections.

4.1. Robust Position Patch

The RPP method proposed in this work extends the original contribution of PP [6] and assumes that low- and high-quality manifolds have similar local structure. This method derives the combination weights $\mathbf{w}_i$ by solving

$$\mathbf{w}_i = \arg\min_{\mathbf{w}_i} ||\widehat{\mathbf{x}}_i - \widehat{\mathbf{L}}_i \mathbf{w}_i||_2^2 \text{ subject to } ||\mathbf{w}_i||_2^2 = 1 \quad (3)$$

The above formulation tries to find the optimal combination weights on the refined low-resolution dictionary $\widehat{\mathbf{L}}_i$ which best represent the low-quality test patch $\widehat{\mathbf{x}}_i$ and has a closed form solution. The same reconstruction weights $\mathbf{w}_i$ are then used to reconstruct the i-th patch using

$$\widetilde{\mathbf{y}}_i^{\text{RPP}} = \mathbf{H}_i \mathbf{w}_i \quad (4)$$

4.2. Robust Sparse Position Patch

The RSPP extends the original contribution of SPP [7] and formulates the hallucination problem using

$$\mathbf{u}_i = \arg\min_{\mathbf{u}_i} ||\mathbf{u}_i||_1 \text{ subject to } ||\widehat{\mathbf{x}}_i - \widehat{\mathbf{L}}_i \mathbf{u}_i||_2^2 \leq \epsilon \quad (5)$$

which is a convex problem and can be solved in polynomial time. In this work we use the solver provided by SparseLab[3] to solve the above Basis Pursuit Denoising problem. The hallucinated high-resolution patch is then synthesized using

$$\widetilde{\mathbf{y}}_i^{\text{RSPP}} = \mathbf{H}_i \mathbf{u}_i \quad (6)$$

[3]The code can be found at `https://sparselab.stanford.edu/`

4.3. Robust Multilayer Locality Constrained Iterative Neighbour Embedding (RM-LINE)

The experiments in [18] demonstrated that the manifold assumption on which the above two methods are based does not always hold. The proposed RM-LINE method extends the M-LINE method proposed in [10]. Specifically, the estimate of the high-resolution patch $\widehat{\mathbf{v}}_{0,0}$ is initialized by up-scaling the low-quality patch $\widehat{\mathbf{x}}_i$ using bi-cubic interpolation and the intermediate dictionary $\widehat{\mathbf{L}}_i^{\{0\}} = \widehat{\mathbf{L}}_i$. This iterative method has an outer-loop indexed by $b \in [0, B-1]$ and an inner-loop indexed by $j \in [0, J-1]$. For every iteration of the inner loop, the support $\mathbf{s}$ of $\mathbf{H}_i$ that minimizes the distance

$$\mathbf{d} = ||\widehat{\mathbf{v}}_{j,b} - \mathbf{H}_i(\mathbf{s})||_2^2 \quad (7)$$

is computed using k-nearest neighbours. The combination weights are then derived using

$$\mathbf{w}_{i,j}^\star = \arg\min_{\mathbf{w}_{i,j}^\star} \left(||\widehat{\mathbf{x}}_i - \widehat{\mathbf{L}}_i^{\{b\}}(\mathbf{s}) \mathbf{w}_{i,j}^\star||_2^2 + \tau ||\mathbf{d}(\mathbf{s}) \odot \mathbf{w}_{i,j}^\star||_2^2 \right) \quad (8)$$

where τ is a regularization parameter and $\odot$ denotes the element-wise multiplication. This optimization problem can be solved by an analytic solution [10]. The estimated high-quality patch is then updated using

$$\widehat{\mathbf{v}}_{j+1,b} = \mathbf{H}_i(\mathbf{s}) \mathbf{w}_{i,j}^\star \quad (9)$$

Once all the iterations of the inner-loop are completed, the intermediate dictionary $\widehat{\mathbf{L}}_i^{\{b+1\}}$ is updated using a leave-one-out methodology (see [10] for more detail), and the inner-loop is repeated. The final estimate of the high-resolution patch is then derived using

$$\widetilde{\mathbf{y}}_i^{\text{RM-LINE}} = \widehat{\mathbf{v}}_{J,B-1} \quad (10)$$

5. COMPUTATIONAL COMPLEXITY

The complexity of the proposed method is mainly affected by two parts: i) dictionary construction and ii) the face hallucination method used. Given that the syntax element of a codec has a finite population, the learned dictionaries can be precomputed (one for every syntax element or at least the most commonly used ones) and the syntax element of the image to be enhanced can be used to load the dictionary with the same syntax element. Therefore, the complexity of the dictionary construction can be completely eliminated at the expense of larger storage space requirements. The complexity of the robust face hallucination on the other hand is equivalent to the complexity of the non-robust counterparts, which can be computed in polynomial time.

Table 1: PSNR analysis using H.264/AVC as codec.

Method	d_x	Quantization Parameter		
		20	**25**	**30**
Baseline	10	27.59	27.33	26.63
PP [6]	10	25.74	25.73	25.53
SPP [7]	10	25.52	25.43	25.20
M-LINE [10]	10	25.32	25.22	25.01
RPP	10	**30.19**	**29.73**	**28.65**
RSPP	10	30.12	29.54	28.51
RM-LINE	10	29.95	29.39	28.30
Baseline	20	29.63	29.46	29.03
PP [6]	20	**34.13**	33.29	31.70
SPP [7]	20	33.91	32.89	31.22
M-LINE [10]	20	33.90	32.88	31.25
RPP	20	34.11	**33.36**	**32.05**
RSPP	20	34.01	33.21	31.88
RM-LINE	20	33.94	33.09	31.64

Table 2: Recognition using H.264/AVC Intra as codec.

Method	d_x	Quantization Parameter		
		20	**25**	**30**
Baseline	10	0.538	0.498	0.354
PP [6]	10	0.416	0.377	0.281
SPP [7]	10	0.469	0.399	0.304
M-LINE [10]	10	0.485	0.424	0.293
RPP	10	0.553	0.486	0.318
RSPP	10	**0.619**	**0.561**	0.385
RM-LINE	10	0.607	0.556	**0.394**
Baseline	20	0.722	0.693	0.625
PP [6]	20	0.748	0.714	0.640
SPP [7]	20	**0.780**	0.732	0.667
M-LINE [10]	20	0.779	**0.744**	**0.681**
RPP	20	0.745	0.700	0.623
RSPP	20	0.761	0.723	0.634
RM-LINE	20	0.770	0.724	0.654

6. EXPERIMENTAL SET-UP

The set $\mathcal{H}$ includes images from both Color Feret [19] and Multi-Pie [20] datasets, where only frontal facial images were considered. One image per subject was randomly selected, resulting in a dictionary size $m = 1203$. The gallery combined facial image (one image per subject) from the FRGC-V2 [21] dataset (controlled environment) and MEDS-II [22] datasets, providing a gallery of 889 face images. The probe images were taken from the FRGC-V2 (uncontrolled environment), where two images per subject were included, resulting in a probe set of 930 images. All images were registered such that $d_y = 40$. The face hallucination methods were configured such that n = 25 and γ_x = 2, while the algorithm specific parameters of both original and extended face hallucination methods considered here were set as specified in the respective papers. The face recognition analysis was conducted using the LBP face recognizer [23]. Due to limit of space, only part of the experimental results are reported in this paper[4].

7. SIMULATION RESULTS

The quality analysis in terms of PSNR of H.264/AVC Intra compressed images are summarized in table 1. These results were generated using the JM version 19.0 to simulate H.264/AVC compression. The compression of the H.264/AVC Intra (Baseline profile) was controlled using the Quantization Parameter where the de-blocking filter was enabled. These results clearly show that the proposed RPP, RSPP and RM-LINE methods outperform the baseline (H.264/AVC decoded followed by bi-cubic upscaled image) and the face super-resolution methods presented in [6, 7, 10] which are not robust to the distortions caused by compression. More specifically, RPP was found to provide the best performance achieving PSNR gains higher than 2dB for H.264/AVC encoded face images.

The recognition performance for H.264/AVC restored images is summarized in table 2 where it can be seen that the proposed methods manage to improve the recognition performance over the baseline, achieving rank-1 recognition improvement between 2.9 – 8.1 %. It can also be seen that the proposed methods (especially RSPP and RM-LINE) achieve recognition rates up to 16.2% at $d_x = 10$ compared to SPP and M-LINE respectively while being competitive to super-resolution methods at $d_x = 20$. The reduction in recognition at $d_x = 20$ can be explained by the fact that since our method performs up-scaling and de-noising jointly, the de-noising part will blur the restored image thus reducing the texture detail crucial for recognition.

8. COMMENTS AND CONCLUSION

This work presents a new dictionary construction method which employs the syntax available in the image file to model the distortions afflicting the low-quality image $\widehat{\mathbf{X}}$. This method generates low-quality dictionaries, whose noise characteristics are more representative of $\widehat{\mathbf{X}}$, to make existing face hallucination methods more robust to compression. Experimental results evidence that the dictionaries constructed by the proposed method manage to achieve a significant gain in performance especially at higher magnification factors. However, the robust face hallucination methods presented here are optimized to maximize the PSNR and do not consider recognition in their formulation. Future work points us in the direction to implement face hallucination techniques which incorporate texture consistency within their formulation and extend the proposed method to handle non-frontal face images.

[4]More results can be found at `http://www.reubenfarrugia.com/content/robust-face-hallucination`.

References

[1] M. A. Sasse, "Not seeing the crime for the cameras?," *Commun. ACM*, vol. 53, no. 2, pp. 22–25, Feb. 2010.

[2] J. Yang, H. Tang, Y. Ma, and T. Huang, "Face hallucination via sparse coding," in *IEEE Int. Conf. on Image Processing*, Oct 2008, pp. 1264–1267.

[3] H-Y. Chen and S-Y. Chien, "Eigen-patch: Position-patch based face hallucination using eigen transformation," in *IEEE Int. Conf. on Multimedia and Expo*, July 2014, pp. 1–6.

[4] W. Zhang and W-K. Cham, "Hallucinating face in the dct domain," *IEEE Trans. on Image Processing*, vol. 20, no. 10, pp. 2769–2779, Oct 2011.

[5] B. Kumar and R. Aravind, "Face hallucination using olpp and kernel ridge regression," in *IEEE Int. Conf. on Image Processing*, Oct 2008, pp. 353–356.

[6] X. Ma, J. Zhang, and C. Qi, "Position-based face hallucination method," in *IEEE Int. Conf. on Multimedia and Expo*, June 2009, pp. 290–293.

[7] C. Jung, L. Jiao, B. Liu, and M. Gong, "Position-patch based face hallucination using convex optimization," *IEEE Signal Processing Letters*, vol. 18, no. 6, pp. 367–370, June 2011.

[8] J. Jiang, R. Hu, Z. Han, T. Lu, and K. Huang, "Position-patch based face hallucination via locality-constrained representation," in *IEEE Int. Conf. on Multimedia and Expo*, July 2012, pp. 212–217.

[9] H. Li, L. Xu, and G. Liu, "Face hallucination via similarity constraints," *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 19–22, Jan 2013.

[10] J. Jiang, R. Hu, Z. Wang, and Z. Han, "Face super-resolution via multilayer locality-constrained iterative neighbor embedding and intermediate dictionary learning," *IEEE Trans. on Image Processing*, vol. 23, no. 10, pp. 4220–4231, Oct 2014.

[11] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, "A comprehensive survey to face hallucination," *Int. J. of Computer Vision*, vol. 106, no. 1, pp. 9–30, 2014.

[12] B. K. Gunturk, Y. Altunbasak, and R.M. Mersereau, "Super-resolution reconstruction of compressed video using transform-domain statistics," *IEEE Trans. on Image Processing*, vol. 13, no. 1, pp. 33–43, Jan 2004.

[13] C. A. Segall, R. Molina, and A.K. Katsaggelos, "High-resolution images from low-resolution compressed video," *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 37–48, May 2003.

[14] D. Sun and W-K. Cham, "Postprocessing of low bit-rate block dct coded images based on a fields of experts prior," *IEEE Trans. on Image Processing*, vol. 16, no. 11, pp. 2743–2751, Nov 2007.

[15] X. Zhang, R. Xiong, X. Fan, S. Ma, and W. Gao, "Compression artifact reduction by overlapped-block transform coefficient estimation with block similarity," *IEEE Trans. on Image Processing*, vol. 22, no. 12, pp. 4613–4626, Dec 2013.

[16] J. Zhang, S. Ma, Y. Zhang, and W. Gao, "Image deblocking using group-based sparse representation and quantization constraint prior," in *IEEE Int. Conf. on Image Processing*, Sept 2015, pp. 306–310.

[17] A. Nosratinia, "Postprocessing of jpeg-2000 images to remove compression artifacts," *IEEE Signal Processing Letters*, vol. 10, no. 10, pp. 296–299, Oct 2003.

[18] B. Li, H. Chang, S. Shan, and X. Chen, "Locality preserving constraints for super-resolution with neighbor embedding," in *IEEE Int. Conf. on Image Processing*, Nov 2009, pp. 1189–1192.

[19] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295–306, Apr. 1998.

[20] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image Vision Comput.*, vol. 28, no. 5, pp. 807–813, May 2010.

[21] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2005, pp. 947–954.

[22] A. P. Founds, N. Orlans, G. Whiddon, and C. Watson, "Nist special database 32 multiple encounter dataset ii (meda-ii)," Tech. Rep., National Institute of Standards and Technology, 2011.

[23] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.