Deep Face Deblurring

Grigorios G. Chrysos
Imperial College London
g.chrysos@imperial.ac.uk

Stefanos Zafeiriou
Imperial College London
s.zafeiriou@imperial.ac.uk

Abstract

Blind deblurring consists a long studied task, however the outcomes of generic methods are not effective in real world blurred images. Domain-specific methods for deblurring targeted object categories, e.g. *text or faces, frequently outperform their generic counterparts, hence they are attracting an increasing amount of attention. In this work, we develop such a domain-specific method to tackle deblurring of human faces, henceforth referred to as face deblurring. Studying faces is of tremendous significance in computer vision, however face deblurring has yet to demonstrate some convincing results. This can be partly attributed to the combination of i) poor texture and ii) highly structure shape that yield the contour/gradient priors (that are typically used) sub-optimal. In our work instead of making assumptions over the prior, we adopt a learning approach by inserting weak supervision that exploits the well-documented structure of the face. Namely, we utilise a deep network to perform the deblurring and employ a face alignment technique to pre-process each face. We additionally surpass the requirement of the deep network for thousands training samples, by introducing an efficient framework that allows the generation of a large dataset. We utilised this framework to create* $2MF^2$*, a dataset of over two million frames. We conducted experiments with real world blurred facial images and report that our method returns a result close to the sharp natural latent image.*

1. Introduction

Blind deblurring is the task of acquiring an estimate of the sharp latent image given a blurry image as input. No single algorithm for deblurring all objects exists; the task is notoriously ill-posed. To that end, methods that exploit domains-specific knowledge have emerged for deblurring targeted categories of objects, *e.g.* text or faces. Similarly, the focus of this work is face deblurring; we argue that exploiting domain-specific knowledge can lead to superior deblurring results, especially for the human face that presents a highly structured shape. Despite the fact that the human face is among the most studied objects in computer vision with significant applications in face recognition, computer graphics and surveillance, face deblurring has not received much attention yet.

Deblurring has long been studied ([42, 7, 28, 32, 34]), however the results are far from satisfactory ([26]) when it comes to real world blurred images. As illustrated in Fig. 1 the result from state-of-the-art methods in real world blurred images (row 2) is far worse than the synthetically blurred images (row 1). The difficulty in real world blurred images can be attributed to the non-linear functions involved in the imaging process, like lens saturation, depth variation, lossy compression. Nevertheless, optimisation-based deblurring techniques ([33, 28, 17, 34]) have reported some progress, credited to a meticulous choice of priors along with some optimisation restrictions ([28, 35]). Apart from the generic deblurring methods which are applied to all objects ([28, 17, 35]), there are also methods that utilise domain-specific knowledge, *e.g.* text or face priors ([33, 32]). Domain-specific methods frequently outperform their generic counterparts due to their stronger form of supervision.

The human face includes some characteristics, *e.g.* fairly restricted shape, that allow a stronger form of supervision. To the best of our knowledge, the method of [32] is currently the only method that explicitly models the blurring for the human face. The authors' motivation relies in capitalising on the restricted facial shape to guide their optimisation. Their method computes the external contour of the face and matches it with an exemplar image; then the contour of the exemplar match is used as a prior. The contour matching restricts the usage of the method since a) it is computationally demanding to compare each image against a dataset, b) the matching is inaccurate for poses that do not exist in the dataset. In contrast to [32], most of the generic methods yield sub-optimal results in face deblurring, since they include either a prior based on the gradient or a contour/edge detection step. The highly structured facial shape along with the poor texture constitute the reasons why those generic methods are sub-optimal.

In our work, instead of 'intuitively' adding priors or

(a) Blurred image (b) Babacan *et al*. [2] (c) Zhang *et al*. [45] (d) Pan *et al*. [32] (e) Pan *et al*. [33] (f) Pan *et al*. [34] (g) Original image

Figure 1: (Preferably viewed in colour) Two sample facial images as deblurred by the existing methods. The one on the top row was synthetically blurred with a uniform kernel, while the one on the bottom is a real world blurred image. Evidently, the existing methods do not yield a sharp natural facial image as we would expect. The difference between deblurring results depending on the type of blur as emphasized in [26] can be visually confirmed.

making other assumptions, we embrace a learning-based approach (by capitalising on the recent developments on Convolutional Neural Networks (CNN)), guided by a weak supervision to express the restriction for the shape structure. The last few years the introduction of elaborate benchmarks [36] allowed CNN methods to surpass the performance of the hand-crafted linear optimisation techniques, *e.g*. in detection [16, 4], model-free tracking [31], classification [20]. The core component of our architecture is the state-of-the-art residual network (ResNet), which is discriminatively trained from training samples of sharp/blurry facial images. A form of weak supervision is introduced by aligning predefined landmark points in the face. This preprocessing step allows the network to encapsulate our restriction for a particular facial structure. We do not enforce a strict alignment as often performed in the landmark localisation techniques ([23, 47]), since warping creates nonlinear artifacts. Additionally, the blurring process might lead to an ambiguity in the exact positioning of the landmarks, hence deblurring might not be as trivial in case of strict alignment. However, in our experimentation the localisation works sufficiently well for our purpose of selecting and pre-processing the region to be fed in the network.

A constraint of the (supervised) learning-based methods is their dependency on massive amounts of training samples. Collecting and annotating such datasets ([36, 6, 46, 39]) is expensive and laborious, hence there is an increasing effort to create datasets semi-automatically [39, 37] or almost in an unsupervised manner [6]. We rectify that for our task by devising an automatic framework that allows the creation of a large dataset with human faces from videos. The framework can select the appropriate frames completely automatically, however in our case a user verified that a face is included in the last frame of each video. We have utilised this framework to create $2MF^2$, a dataset with millions of facial frames. $2MF^2$ consists of over a thousand video clips with an accumulated number of 2,1 million frames, which constitutes $2MF^2$ the largest dataset of video frames for faces[1].

Our contributions can be summarised as:

- We introduce a network architecture that performs face deblurring. We validated the trained model in different experiments including synthetically blurred images, images with simulated motion blur as well as low resolution real world blurred images.
- We introduce an automatic framework that allows the collection of large datasets in a time-efficient manner. We have utilised this framework to create the $2MF^2$ dataset, which consists of more than 2 million frames.

In the following Sections we summarise the related methods in Sec. 2; develop our method in Sec. 3; describe the framework we have devised in Sec. 4 and finally experimentally validate our method in Sec. 5.

2. Related Work

Blur is typically modelled as the convolution of a blur kernel $\boldsymbol{K}_g$ with a (latent) sharp image $\boldsymbol{I}$, i.e.

$$\boldsymbol{I}_{bl} = \psi(\boldsymbol{I} * \boldsymbol{K}_g + \boldsymbol{\epsilon}) \tag{1}$$

with $\boldsymbol{I} \in \mathbb{R}^{h_1 \times w_1}$, $\boldsymbol{K}_g \in \mathbb{R}^{h_2 \times w_2}$ ($h_2 \ll h_1$, $w_2 \ll w_1$), while $\boldsymbol{I}_{bl} \in \mathbb{R}^{a \times b}$ denotes the blurry image with $a = h_1 - h_2 + dh, b = w_1 - w_2 + dw$. The dh, dw depend on the type

[1]The alternatives of 300VW [39] and Youtube Faces [44] include 250 and 620 thousand frames respectively. Furthermore, the Youtube Faces is not appropriate for discriminative learning, since many of the clips are already blurred and of low resolution.

of convolution ('$*$') chosen. The symbol of $\boldsymbol{\epsilon}$ represents the noise term, ψ a function that models additional non-linear artifacts, *e.g*. lossy compression, saturated regions, non-linear sensor response. Retrieving the sharp image is an ill-posed problem, thus some strong assumptions/priors are required. A simple illustration of the ill-posed nature of the task is that for any fixed solution $\tilde{\boldsymbol{I}}, \tilde{\boldsymbol{K}}_g$ of Eq. 1, the family of $\tilde{\boldsymbol{I}} \cdot \lambda, \frac{\tilde{\boldsymbol{K}}_g}{\lambda}$ is also a valid solution, which is referred to as the scaling ambiguity.

Blind deblurring methods can be divided into three categories, based on the approach for obtaining the sharp image (estimation of the latent image in Eq. 1). Each category includes an extensive literature, hence only the most closely related to our work are summarised below. For a more thorough study, the interested reader is redirected to [26].

Synthesis-based: Instead of solving the optimisation, these methods typically include a heuristic for 'guessing' the blurry parts and then apply a synthesis-based replacement in the blurry regions. The majority of those works ([8, 41]) implicitly assume that i) there are multiple frames with approximately the same content and that ii) there exists a sharp patch that matches in content the respective blurry one. These two strong assumptions combined with the poor texture in a face (which weakens the heuristic to detect sharp patches), result in not using synthesis-based methods for face deblurring.

Optimisation-based: Based on Eq. 1 and assuming ψ is the identity function, this class of methods formulates the problem as the minimisation of a cost function of the format

$$\tilde{\boldsymbol{I}} = \underset{\boldsymbol{I}}{\operatorname{argmin}}(||\boldsymbol{I}_{bl} - \boldsymbol{I} * \boldsymbol{K}_g||_2^2 + f(\boldsymbol{I}_{bl})). \qquad (2)$$

with $f(\boldsymbol{I}_{bl})$ a set of priors based on generic image statistics or domain-specific priors. These methods are applied in a coarse-to-fine manner, while they estimate the (dense) kernel and then perform a non-blind deconvolution.

The estimation of the blur kernel $\boldsymbol{K}_g$ and the latent image $\boldsymbol{I}$ occur in an alternating manner, which might lead to a blurry $\tilde{\boldsymbol{I}}$ if a joint MAP (Maximum a posteriori) optimisation is followed ([28]). Levin *et al*. suggest instead to solve a MAP on the kernel with a gradient-based prior based natural image statistics. More recently, Pan *et al*. in [33] apply an ℓ_0 norm as a sparse prior on both the intensity values and the image gradient for deblurring text. HaCohen *et al*. in [17] support that the gradient prior alone is not sufficient, and introduce a prior that locates dense correspondences of the blurry image with a similar sharp image, while they iteratively optimise over the correspondence, the kernel and the sharp image estimation. A strong requirement of their algorithm is the similar reference image, which is not always available. A generalisation of [17] is the work of [32], which also requires an exemplar dataset to locate an image with a similar contour. However, in [32] the authors restrict the task to face deblurring to profit from the shape structure. A search in a dataset with exemplar images is performed to locate an image with a similar contour as the test image. The gradient of the exemplar image provides the initial blind estimation iterations, which leads to an improved performance. Unfortunately, the noisy contour matching process along with the obligatory presence of a similar contour in the dataset limit the applications of this work.

Even though the optimisation-based methods have proven to work well with synthetic blurs, they do not generalise well in real world blurred images ([26]) due to the strong assumptions of invariance and the simplified format of ψ. Another common attribute of these methods is the iterative optimisation procedure; they are executed in a loop hundreds or even thousands of times to return a deblurred image, which classifies these methods as computationally intensive; some of them require hours for deblurring a single image ([5]).

Learning-based: With the resurrection of neural networks, few approaches for learning a network to perform deblurring have emerged. The experimental superiority of neural networks as function approximators consists a strong motivation for relying on neural networks for deblurring. The non-linear units allow us to model non-linear functions ψ or spatially varying blur kernels. Obtaining a sharp image in this case is defined as a function $\tilde{\boldsymbol{I}} = \phi(\boldsymbol{I}_{bl}, \boldsymbol{p})$, with $\boldsymbol{p}$ denoting the hyper-parameters of the method.

Some methods ([21]) learn straight away the function ϕ from the data, while others ([40, 5]) learn an estimate and perform non-blind de-blurring/refinement of the sharp image. In [21], the regularised ℓ_2 loss of an up to 15-layer CNN is minimised for text deblurring. Even though they report nice results, the text deblurring domain is a structured but limited class (the sharp text can be represented as a sequence of binary intensity values). They argue that the performance can be mainly attributed to the network that modelled well the text prior, hence it is questionable whether this would work in more complex object types. Sun *et al*. in [40], learn a CNN to recognise few discretised motion kernels and then perform a non-blind deconvolution in a dense motion field estimate.

Our method belongs in the learning-based category, specifically the methods that learn ϕ from the data. The combination of such a learning method with weak supervision through landmark localisation has not been performed before for deblurring.

3. Method

In this Section, we portray our learning-based method for face deblurring. We develop our way for providing the required input for the network (pairs of blurry/sharp images). Sequentially we introduce the deep architecture that we employed, along with the pre-processing step to take advan-

tage of the facial structure through landmark localisation. Finally, we refer to the inference steps for an unseen image.

3.1. Notation

A sparse shape of n fiducial (landmark) points is denoted as $\boldsymbol{l}$ for the image $\boldsymbol{I}$ with $\boldsymbol{l} = [[\boldsymbol{\ell}_1]^T, [\boldsymbol{\ell}_2]^T, ..., [\boldsymbol{\ell}_n]^T]^T$, with $\boldsymbol{\ell}_j = [x_j, y_j]^T, j \in [1, n], x_j, y_j \in \mathbb{R}$ the Cartesian coordinates of the j[th] point. When referring to a random image $\boldsymbol{I}$, we hypothesise that $\boldsymbol{I}$ contains a human face, of which the facial sparse shape $\boldsymbol{l}$ is available.

3.2. Training pair creation

The dominant way to discriminatively train a network is by feeding pairs of input and label samples; the labels are used to compute the error and improve the network performance. In our case the input is the blurry image, the label is the corresponding sharp image. Obtaining real world blurred images with a dense correspondence with a similar sharp image is not trivial, especially if thousands such pairs are required to train a deep network. Hence, following similar methods ([40, 5]) we resort to simulating the blur from sharp images.

A synthetically blurred image $\boldsymbol{I}_{bl}$ is generated by convolving the original sharp image $\boldsymbol{I}$ with a blur kernel (simulating Eq. 1). A unique blur kernel is created for every input image to allow for the maximum variation in the number of blur kernels that have emerged during the training. The blur kernel is chosen arbitrarily in each step between a Gaussian blur kernel and a motion blur kernel, both with varying deviation and spatial support.

3.3. Network

Following the reasoning of [21] that demonstrated the success of a 15-layer CNN for the simpler task of text deblurring, we employ a network with several convolutional layers to allow a richer representation to be learnt. A modified version of the residual network (ResNet) architecture of [20] is used as the learning component in our method. ResNet consists of a number of 'blocks'; each 'block' is a sequence of convolutional layers, followed by Rectified Linear Units, with identity connections connecting the blocks. This simple architecture has demonstrated state-of-the-art performance in several tasks, while there is an effort to establish their dominance from a theoretical perspective ([18]).

We modify the original ResNet by disabling all the max pooling operations, while skip connections ([19]) are added in the 2^{nd} and 3^{rd} ResNet blocks. A batch normalisation is added in every skip connection to ensure a common scale; a linear mapping is learnt from the high-dimensional space of the connections to the low dimensional space of the output image shape. The huber loss of [22] is utilised for our loss function. This is a continuous and differentiable function with

$$L_h(\boldsymbol{x}) = \begin{cases} ||\boldsymbol{x}||_1 - 0.5 & ||\boldsymbol{x}||_1 > 1 \\ 0.5||\boldsymbol{x}||_2^2 & otherwise \end{cases} \quad (3)$$

Namely, the loss function of our network is:

$$L = L_h(\phi(\boldsymbol{I}_{bl}) - \boldsymbol{I}) \quad (4)$$

3.4. Inference

The single input during inference is the blurry image $\boldsymbol{I}_{bl}$, *i.e*. no latent image as ground-truth is required. The blurry image is pre-processed to obtain the appropriate region of the image to be fed into the network. Concretely, an off-the-shelf face detector is employed to acquire the bounding box; the landmarks are localised through a localisation technique. The image is rescaled based on the size of the landmarks, while a rectangular area around the face (landmarks) is cropped, which is the area that is fed into the network (only the feed-forward part of which is required).

Among the most successful face detectors is the deformable part models (DPM) detector [15, 30]. DPM learn a mixture of models which aim to detect faces in different poses. Each model implicitly considers some parts which are allowed to deform with a quadratic cost. The cost function of DPM contains an appearance (unary) term along with a pairwise (deformation) term plus a bias, all of which are learned with a discriminative training procedure. The crude bounding box of the DPM consists the initialisation of a landmark localisation technique [47, 23, 9]. Both techniques [47, 23] belong to the regression based discriminative methods for landmark localisation. These methods learn to regress from the pixel intensities (with the former extracting hand-crafted SIFT features, while the latter of Kazemi *et al*. rely on data driven learned features) to the sparse shape coordinates. Both methods have proven very accurate in a number of benchmarks [37, 9], hence we adopt the method of Kazemi *et al*. due to a publicly available fast implementation [24].

4. Data mining

In this Section, we describe our method for mining frames from videos in a semi-supervised manner. A number of videos are crawled using the API's of web sources, *e.g*. Youtube; each video consists of few thousand frames and is analysed independently to determine the frames, if any, that are appropriate for the task. In our case, we aim at utilising the videos with dynamically moving faces. We defined the following three requirements for a video to be included in the training:

1. a face is present in each frame,

2. the face is not completely static throughout the video,

Figure 2: Sample frames from the $2MF^2$ dataset along with the sparse shape denoted with white dots.

3. the video includes real world images, not synthetically generated ones.

To that end, we have devised an efficient, automatic framework to perform this task; the steps are summarised in Alg. 1.

The face detector of [15, 30] is applied to the first frame of the video. If there is no detection, the video is discarded, otherwise the bounding box obtained initialises a model-free tracker. Given the state of the first frame, a model-free tracker determines the state of the subsequent frames, while no prior information about the object is provided. The tracker should adapt to any appearance, deformation changes, which constitutes a very challenging task, thus an immense amount of diverse techniques has been proposed. In our work, we utilise the SRDCF tracker [11], which provides a decent trade-off of accurate deformable tracking quality and computation complexity [9].

Even though SRDCF is robust to a wide range of variations, an additional criterion of overlap per frame with the bounding box of the DPM detector is performed. Specifically, we require the bounding boxes of the tracker and the detector to have at least a 50% overlap (intersection over union overlap) in half of the frames, otherwise the clip is discarded. Subsequently, the landmark localisation technique of [47] is employed to obtain the sparse shape for each face. Due to the object-agnostic nature of the model-free tracker, we eliminate the few erroneous fittings by learning a statistical function f_{cl}. We utilise a linear patch-based SVM [10] as the classifier $f_{cl}(\boldsymbol{I}, \boldsymbol{l})$ which accepts a frame $\boldsymbol{I}$ along with the respective fitting $\boldsymbol{l}$ and returns a binary decision on whether this is an acceptable fitting. The classifier fulfils the first requirement for every frame, *i.e.* that a face is present.

The requirement of non-static faces is fulfilled by computing the optical flow [14] in the accepted frames and requiring that there is at least a pixel movement from frame to frame. If the average movement per pixel is above a threshold, the video is discarded.

This framework can be adapted for different type of objects with two minor modifications. The modifications are: (i) the face detection module, which can be trivially replaced by a generic detector like [16], (ii) the classifier module for the removal of erroneous fittings, which should be trained for the task, *e.g.* to accept the whole bounding box instead of the patch-based SVM utilising the landmarks.

Algorithm 1: The automatic framework as introduced in Sec. 4 to create the $2MF^2$ dataset.

Input : Video frames $\boldsymbol{V} = [\boldsymbol{I}^{(1)}, \boldsymbol{I}^{(2)}, \ldots, \boldsymbol{I}^{(M)}]$
Output : Accepted frames $\boldsymbol{F}$, Landmarks $\boldsymbol{L}$
Initialize: $\boldsymbol{F} = [], \boldsymbol{L} = [], cnt\_over = 0$
`/* detection in the first frame. */`
$\boldsymbol{faces}$ = face-detection($\boldsymbol{I}^{(1)}$)
if *length*($\boldsymbol{faces}$) $== 0$ **then**
  return $\boldsymbol{F}, \boldsymbol{L}$
end
`/* bb: tracker's bounding box. */`
$\boldsymbol{bb} = \boldsymbol{faces}[0]$
`/* main tracking loop. */`
for *idx = 1 to M* **do**
  $\boldsymbol{faces}$ = face-detection($\boldsymbol{I}^{(idx)}$)
  $\boldsymbol{bb}$ = track($\boldsymbol{I}^{(idx)}$, $\boldsymbol{bb}$)
  if *length*($\boldsymbol{faces}$) > 0 *and* *compute_overlap*($\boldsymbol{faces}[0]$, $\boldsymbol{bb}$) > 0.5 **then**
    $cnt\_over += 1$
  end
  $\boldsymbol{l}^{(idx)} = landmark\_localisation(\boldsymbol{I}^{(idx)}, \boldsymbol{bb})$
  `/* ` f_{cl}`: classifier to reject the erroneous fittings. */`
  accept_fitting = $f_{cl}(\boldsymbol{I}^{(idx)}, \boldsymbol{l}^{(idx)})$
  if *accept_fitting* **then**
    append($\boldsymbol{F}, \boldsymbol{I}^{(idx)}$)
    append($\boldsymbol{L}, \boldsymbol{l}^{(idx)}$)
  end
end
if $cnt\_over < M/2$ **then**
  return [], []
end
return $\boldsymbol{F}, \boldsymbol{L}$

The aforementioned framework was utilised to create $2MF^2$ (2 million frames of faces). $2MF^2$ consists of 1150 videos, with 2,1 million accepted frames that contain a human face. Exemplar frames of few videos are visualised

in Fig. 2, while an accompanying video depicting accepted frames along with the derived sparse shape can be found in https://youtu.be/Mz0918XdDew.

5. Experiments

In this Section we develop few implementation details, summarise a validation experiment for our method with a simple Gaussian blur, compare with the state-of-the-art methods for deblurring in two different scenarios, which include motion blur and real world blurred images.

5.1. Implementation details

The network was implemented in Tensorflow [12] using the Python API; the pre-trained weights of the network were obtained from the original ResNet paper [20], while the majority of the rest functionality was provided by the Menpo project [1].

Figure 3: (Preferably viewed in colour) Visual results for the self evaluation experiment. On the first row is the original image, on the middle the blurred image and the last consists of the outputs of our network.

The shapes of the public datasets with 68 facial points mark-up annotation, *i.e.* IBUG [38], HELEN [27], LFPW [3] and the 300W [37] were utilised a) for training the classifier of Sec.4, b) as additional input to the network for training. Few images with severe distortions were excluded from the training set; the frames of $2MF^2$ were subsampled and one every 2^{nd} frame was used for the training.

The training steps of the classifier were the following: (a) The positive training samples were extracted from the 300W trainset; perturbed versions of the annotations of those images along with selected images of Pascal dataset [13] were used for mining the negative samples. (b) A fixed size patch was extracted from each positive sample around each of the n landmark points; SIFT [29] were computed per patch. For each negative sample a random perturbation of the ground truth points was performed to create an erroneous fitting prior to extracting the patches. (c) A linear SVM was trained, with its hyper-parameters cross-validated in withheld validation frames.

Image type/Quality metric	PSNR	SSIM
Blurred	21.84	0.52
Deblurred	22.42	0.57

Table 1: Image quality metrics for the validation of the network's outputs.

For training our network, we used a mini-batch size of 16; SGD with an exponentially decreasing learning rate (initial value of 0.0003), and decreasing by a factor of 0.5 every 15k iterations. The final training consisted of 70k iterations and was completed in a single-core GPU machine. It should be noted that each frame was loaded only once in the network, to avoid over-fitting the training data. Our method functions at 6 fps in a GPU Titan X machine.

5.2. Self evaluation

Seventy images of AFLW [25] were used to validate the outcome of the network. The images were synthetically blurred with Gaussian noise, while the standard visual quality metrics of PSNR and SSIM [43] were employed to compare the blurred images with the outputs of our network. The quality metrics are reported in Tab. 1 and few indicative images are visualised in Fig. 3. Both the qualitative and quantiative metrics indicate that the method indeed works well under Gaussian blur.

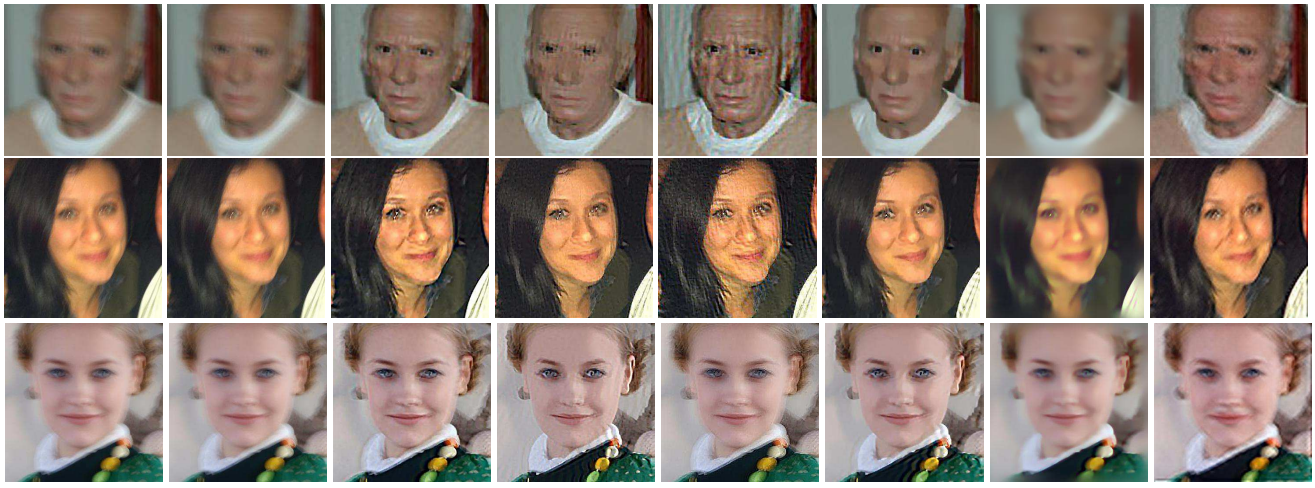
5.3. Simulated motion blur

To extend the simple Gaussian blur, an experiment that simulates motion blur observed in real world blurred images was conducted (this process of simulating the motion blur was only performed during testing). A set of sequential frames of a high frame rate video are averaged and simulate the movement of the person. The averaging creates the effect of a dynamic movement, while the middle frame of the averaging can be considered as the ground-truth frame. The edge cases of this simulation consist of (a) no movement case, (b) extreme movement case. The former case was avoided by considering the optical flow of each two sequential frames and ensuring there is at least some movement in the scene from frame to frame. For the latter case, the PSNR of the averaged frame was compared against the middle frame (ground-truth) and the frames below a threshold were discarded as too noisy.

In our experiment, four videos of the 300VW dataset [39] were utilised. All the videos of 300VW include a single person per video, while they are all over 25 fps.

(a) GT (b) Blurred (c) [2] (d) [45] (e) [32] (f) [33] (g) [34] (h) [5] (i) Proposed

Figure 4: (Preferably viewed in colour) Qualitative results for the simulated blur experiment.



(a) Blurred image (b) [2] (c) [45] (d) [33] (e) [32] (f) [34] (g) [5] (h) Proposed

Figure 5: (Preferably viewed in colour) Some images from the dataset of Lai *et al*. [26]. Notice that our method avoids the over-smoothing of other methods, *e.g*. [34]. Even though it deblurs the texture of the skin in a decent way, it sometimes suffers in localising the iris of the eye.

For each one of the videos employed, a different number of frames were averaged, ranging from 7 to 11 sequential frames. Also, the recent deblurring methods of Babacan *et al*. [2], Zhang *et al*. [45], Pan *et al*. [32], Pan *et al*. [33], Pan *et al*. [34] and Chakrabarti [5] were also included in the experiment. In Fig. 4, the qualitative results of frames with simulated blur are visualised, while in Tab. 2 the quantitative metrics are reported.

5.4. Real world blurred images

Providing a method that works for real world blurred images consists a strong motivation for our work. Unfortunately, comparing with real world blurred images comes at the cost of not having any ground-truth image[2]. Therefore,

[2]Capturing a real world blurred image and a sharp one with a dense correspondence requires an elaborate hardware/software setup. An approximation can be considered by capturing videos with a high frame rate camera (the middle frame can be used as the ground-truth), however this still does not guarantee the simulation to real world blurred image.

Figure 6: (Preferably viewed in colour) Qualitative results in real world blurred images from arbitrary videos. On the top row, the original frame (there is no ground-truth available); on the second row the output of our method.

Image type/Quality metric	PSNR	SSIM
Babacan *et al*. [2]	25.127	0.580
Zhang *et al*. [45]	23.303	0.521
Pan *et al*. [32]	21.304	0.476
Pan *et al*. [33]	22.492	0.473
Pan *et al*. [34]	23.972	0.512
Chakrabarti [5]	23.388	0.420
Proposed	23.950	0.558

Table 2: Image quality metrics for the simulated motion blur experiment of Sec. 5.3.

we opted to report the visual comparisons here.

In Fig. 5, the comparisons among different methods are provided for the facial images of Lai *et al*. [26]. Additionally, to further emphasise the merits of the proposed method, we have gathered few images from internet sources in both indoors and outdoors scenes. The faces in those frames are of quite low-resolution, while there is rapid movement in the scne. The qualitative results are visualised in Fig. 6.

6. Discussion and conclusions

In this work, we introduced a new method for deblurring facial images through inserting a weak supervision in the system, but not explicitly enforcing a strict alignment. The architecture that we have implemented is a modified version of the strong performing ResNet. We have also developed an automatic framework for large dataset creation with off-the-shelf tools from the literature. Moreover, we have created $2MF^2$, a dataset that includes more than one thousand clips containing over two million frames of faces. The dataset was utilised to perform the training of our network. A number of experiments are conducted to validate the performance of our method and compare against the state-of-the-art deblurring methods.

7. Acknowledgements

G. Chrysos was supported by EPSRC DTA award at Imperial College London. S. Zafeiriou was partially funded by the EPSRC Project EP/N007743/1 (FACER2VM).

References

[1] J. Alabort-i-Medina, E. Antonakos, J. Booth, P. Snape, and S. Zafeiriou. Menpo: A comprehensive platform for parametric image alignment and visual deformable models. In *Proceedings of ACM International Conference on Multimedia (ACM'MM)*, pages 679–682. ACM, 2014. [Code: http://www.menpo.org/, Status: Online; accessed 9-November-2016]. 6

[2] S. D. Babacan, R. Molina, M. N. Do, and A. K. Katsaggelos. Bayesian blind deconvolution with general sparse image priors. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 341–355. Springer, 2012. 2, 7, 8

[3] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 35(12):2930–2940, 2013. 6

[4] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick. Inside-outside net: Detecting objects in context with skip

pooling and recurrent neural networks. *arXiv preprint arXiv:1512.04143*, 2015. 2

[5] A. Chakrabarti. A neural approach to blind motion deblurring. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 221–235. Springer, 2016. 3, 4, 7, 8

[6] J. Charles, T. Pfister, M. Everingham, and A. Zisserman. Automatic and efficient human pose estimation for sign language videos. *International Journal of Computer Vision (IJCV)*, 110(1):70–90, 2014. [Data: https://www.robots.ox.ac.uk/~vgg/data/pose/, Status: Online; accessed 9-November-2016]. 2

[7] C.-M. Cho and H.-S. Don. Blur identification and image restoration using a multilayer neural network. In *Neural Networks, 1991. 1991 IEEE International Joint Conference on*, pages 2558–2563. IEEE, 1991. 1

[8] S. Cho, J. Wang, and S. Lee. Video deblurring for hand-held cameras using patch-based synthesis. *ACM Trans. Gr.*, 31(4):64, 2012. 3

[9] G. G. Chrysos, E. Antonakos, P. Snape, A. Asthana, and S. Zafeiriou. A comprehensive performance evaluation of deformable face tracking "in-the-wild". *International Journal of Computer Vision (IJCV)*, 2017. 4, 5

[10] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 5

[11] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg. Learning spatially regularized correlation filters for visual tracking. In *IEEE Proceedings of International Conference on Computer Vision (ICCV)*, pages 4310–4318, 2015. [Code: https://www.cvl.isy.liu.se/en/research/objrec/visualtracking/regvistrack/, Status: Online; accessed 9-November-2016]. 5

[12] M. A. *et al*. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. [Code: http://tensorflow.org/, Status: Online; accessed 9-November-2016]. 6

[13] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 88(2):303–338, 2010. [Data: http://host.robots.ox.ac.uk/pascal/VOC/voc2007/index.html, Status: Online; accessed 15-January-2017]. 6

[14] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. *Image analysis*, pages 363–370, 2003. 5

[15] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 32(9):1627–1645, 2010. 4, 5

[16] R. Girshick. Fast r-cnn. In *IEEE Proceedings of International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. 2, 5

[17] Y. Hacohen, E. Shechtman, and D. Lischinski. Deblurring by example using dense correspondence. In *IEEE Proceedings of International Conference on Computer Vision (ICCV)*, pages 2384–2391, 2013. 1, 3

[18] M. Hardt and T. Ma. Identity matters in deep learning. *ICLR*, 2017. 4

[19] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 447–456, 2015. 4

[20] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016. 2, 4, 6

[21] M. Hradiš, J. Kotera, P. Zemcík, and F. Šroubek. Convolutional neural networks for direct text deblurring. In *Proceedings of British Machine Vision Conference (BMVC)*, 2015. 3, 4

[22] P. J. Huber. Robust regression: asymptotics, conjectures and monte carlo. *The Annals of Statistics*, pages 799–821, 1973. 4

[23] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1867–1874, 2014. 2, 4

[24] D. E. King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009. [Code: http://dlib.net/, Status: Online; accessed 9-November-2016]. 4

[25] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2144–2151. IEEE, 2011. 6

[26] W.-S. Lai, J.-B. Huang, Z. Hu, N. Ahuja, and M.-H. Yang. A comparative study for single image blind deblurring. In *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016. 1, 2, 3, 7, 8

[27] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 679–692. Springer, 2012. 6

[28] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman. Understanding and evaluating blind deconvolution algorithms. In *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1964–1971. IEEE, 2009. 1, 3

[29] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004. 6

[30] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 720–735. Springer, 2014. 4, 5

[31] H. Nam and B. Han. Learning multi-domain convolutional neural networks for visual tracking. In *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016. 2

[32] J. Pan, Z. Hu, Z. Su, and M.-H. Yang. Deblurring face images with exemplars. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 47–62. Springer, 2014. 1, 2, 3, 7, 8

[33] J. Pan, Z. Hu, Z. Su, and M.-H. Yang. Deblurring text images via l0-regularized intensity and gradient prior. In *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2901–2908, 2014. 1, 2, 3, 7, 8

[34] J. Pan, D. Sun, H. Pfister, and M.-H. Yang. Blind image deblurring using dark channel prior. In *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016. 1, 2, 7, 8

[35] D. Perrone and P. Favaro. Total variation blind deconvolution: The devil is in the details. In *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2909–2916. IEEE, 2014. 1

[36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 2

[37] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. In *Image and Vision Computing*, 2015. 2, 4, 6

[38] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *IEEE Proceedings of International Conference on Computer Vision (ICCV-W), 300 Faces In-the-Wild Challenge (300-W)*, pages 397–403, 2013. 6

[39] J. Shen, S. Zafeiriou, G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *IEEE Proceedings of International Conference on Computer Vision, 300 Videos in the Wild (300-VW): Facial Landmark Tracking in-the-Wild Challenge & Workshop (ICCV-W)*, December 2015. [Data: http://ibug.doc.ic.ac.uk/resources/300-VW/, Status: Online; accessed 9-November-2016]. 2, 6

[40] J. Sun, W. Cao, Z. Xu, and J. Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 769–777. IEEE, 2015. 3, 4

[41] F. Tan, S. Liu, L. Zeng, and B. Zeng. Kernel-free video deblurring via synthesis. In *IEEE Proceedings of International Conference on Image Processing (ICIP)*, pages 2683–2687. IEEE, 2016. 3

[42] A. Tekalp, H. Kaufman, and J. Woods. Identification of image and blur parameters for the restoration of noncausal blurs. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(4):963–972, 1986. 1

[43] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions in Image Processing (TIP)*, 13(4):600–612, 2004. 6

[44] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 529–534. IEEE, 2011. 2

[45] H. Zhang, D. Wipf, and Y. Zhang. Multi-image blind deblurring using a coupled adaptive sparse prior. In *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1051–1058. IEEE, 2013. 2, 7, 8

[46] W. Zhang, M. Zhu, and K. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2248–2255, 2013. [Data: http://dreamdragon.github.io/PennAction/, Status: Online; accessed 9-November-2016]. 2

[47] S. Zhu, C. Li, C. Change Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4998–5006, 2015. 2, 4, 5