Single image super-resolution based on neural networks for text and face recognition

Clément Peyrard

N°d'ordre NNT : 2017LYSEI083

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
INSA de Lyon

Ecole Doctorale ED 512
Informatique et Mathématiques de Lyon

Discipline de doctorat :
Informatique

Soutenue publiquement le 29/09/2017, par :
Clément PEYRARD

Single Image Super-Resolution based on Neural Networks for text and face recognition

---

Devant le jury composé de :

M. CHATEAU, Thierry	PRU, Univ. de Clermont-Auvergne	Rapporteur
M. THIRAN, Jean-Philippe	PRU, EPFL	Rapporteur
MME GUILLEMOT, Christine	Directrice de recherche, INRIA	Examinatrice
M. VIARD-GAUDIN, Christian	PRU, Univ. de Nantes	Examinateur
M. GARCIA, Christophe	PRU, INSA de Lyon	Directeur de thèse
M. BACCOUCHE, Moez	Dr, Ingénieur de recherche, Orange	Co-encadrant
M. MAMALET, Franck	Dr, Responsable R&D, Spikenet Technologies	Invité

Département FEDORA – INSA Lyon - Ecoles Doctorales – Quinquennal 2016-2020

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
CHIMIE	**CHIMIE DE LYON** **http://www.edchimie-lyon.fr** Sec : Renée EL MELHEM Bat Blaise Pascal 3e etage secretariat@edchimie-lyon.fr Insa : R. GOURDON	M. Stéphane DANIELE Institut de Recherches sur la Catalyse et l'Environnement de Lyon IRCELYON-UMR 5256 Équipe CDFA 2 avenue Albert Einstein 69626 Villeurbanne cedex directeur@edchimie-lyon.fr
E.E.A.	**ELECTRONIQUE, ELECTROTECHNIQUE, AUTOMATIQUE** **http://edeea.ec-lyon.fr** Sec : M.C. HAVGOUDOUKIAN Ecole-Doctorale.eea@ec-lyon.fr	M. Gérard SCORLETTI Ecole Centrale de Lyon 36 avenue Guy de Collongue 69134 ECULLY Tél : 04.72.18 60.97 Fax : 04 78 43 37 17 Gerard.scorletti@ec-lyon.fr
E2M2	**EVOLUTION, ECOSYSTEME, MICROBIOLOGIE, MODELISATION** **http://e2m2.universite-lyon.fr** Sec : Sylvie ROBERJOT Bât Atrium - UCB Lyon 1 04.72.44.83.62 Insa : H. CHARLES secretariat.e2m2@univ-lyon1.fr	M. Fabrice CORDEY CNRS UMR 5276 Lab. de géologie de Lyon Université Claude Bernard Lyon 1 Bât Géode 2 rue Raphaël Dubois 69622 VILLEURBANNE Cédex Tél : 06.07.53.89.13 cordey@univ-lyon1.fr
EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTE** **http://www.ediss-lyon.fr** Sec : Sylvie ROBERJOT Bât Atrium - UCB Lyon 1 04.72.44.83.62 Insa : M. LAGARDE secretariat.ediss@univ-lyon1.fr	Mme Emmanuelle CANET-SOULAS INSERM U1060, CarMeN lab, Univ. Lyon 1 Bâtiment IMBL 11 avenue Jean Capelle INSA de Lyon 696621 Villeurbanne Tél : 04.72.68.49.09 Fax :04 72 68 49 16 Emmanuelle.canet@univ-lyon1.fr
INFOMATHS	**INFORMATIQUE ET MATHEMATIQUES** **http://infomaths.univ-lyon1.fr** Sec :Renée EL MELHEM Bat Blaise Pascal, 3e étage Tél : 04.72. 43. 80. 46 Fax : 04.72.43.16.87 infomaths@univ-lyon1.fr	M. Luca ZAMBONI Bâtiment Braconnier 43 Boulevard du 11 novembre 1918 69622 VILLEURBANNE Cedex Tél :04 26 23 45 52 zamboni@maths.univ-lyon1.fr
Matériaux	**MATERIAUX DE LYON** **http://ed34.universite-lyon.fr** Sec : Marion COMBE Tél:04-72-43-71-70 –Fax : 87.12 Bat. Direction ed.materiaux@insa-lyon.fr	M. Jean-Yves BUFFIERE INSA de Lyon MATEIS Bâtiment Saint Exupéry 7 avenue Jean Capelle 69621 VILLEURBANNE Cedex Tél : 04.72.43 71.70 Fax 04 72 43 85 28 Ed.materiaux@insa-lyon.fr
MEGA	**MECANIQUE, ENERGETIQUE, GENIE CIVIL, ACOUSTIQUE** **http://mega.universite-lyon.fr** Sec : Marion COMBE Tél:04-72-43-71-70 –Fax : 87.12 Bat. Direction mega@insa-lyon.fr	M. Philippe BOISSE INSA de Lyon Laboratoire LAMCOS Bâtiment Jacquard 25 bis avenue Jean Capelle 69621 VILLEURBANNE Cedex Tél : 04.72 .43.71.70 Fax : 04 72 43 72 37 Philippe.boisse@insa-lyon.fr
ScSo	**ScSo*** **http://recherche.univ-lyon2.fr/scso/** Sec : Viviane POLSINELLI Brigitte DUBOIS Insa : J.Y. TOUSSAINT Tél : 04 78 69 72 76 viviane.polsinelli@univ-lyon2.fr	M. Christian MONTES Université Lyon 2 86 rue Pasteur 69365 LYON Cedex 07 Christian.montes@univ-lyon2.fr

*ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

Abstract

This thesis is focussed on super-resolution (SR) methods for improving automatic recognition system (Optical Character Recognition, face recognition) in realistic contexts.

SR methods allow to generate high resolution images from low resolution ones. Unlike upsampling methods such as interpolation, they restore spatial high frequencies and compensate artefacts such as blur or jaggy edges. In particular, example-based approaches learn and model the relationship between low and high resolution spaces via pairs of low and high resolution images. Artificial Neural Networks are among the most efficient systems to address this problem.

This work demonstrate the interest of SR methods based on neural networks for improved automatic recognition systems. By adapting the data, it is possible to train such Machine Learning algorithms to produce high-resolution images. Convolutional Neural Networks are especially efficient as they are trained to simultaneously extract relevant non-linear features while learning the mapping between low and high resolution spaces.

On document text images, the proposed method improves OCR accuracy by +7.85 points compared with simple interpolation. The creation of an annotated image dataset and the organisation of an international competition (ICDAR2015) highlighted the interest and the relevance of such approaches. Moreover, if a priori knowledge is available, it can be used by a suitable network architecture. For facial images, face features are critical for automatic recognition. A two step method is proposed in which image resolution is first improved, followed by specialised models that focus on the essential features. An off-the-shelf face verification system has its performance improved from +6.91 up to +8.15 points.

Finally, to address the variability of real-world low-resolution images, deep neural networks allow to absorb the diversity of the blurring kernels that characterise the low-resolution images. With a single model, high-resolution images are produced with natural image statistics, without any knowledge of the actual observation model of the low-resolution image.

Résumé

Cette thèse porte sur les méthodes de super-résolution (SR) pour l'amélioration des performances des systèmes de reconnaissance automatique (OCR, reconnaissance faciale).

Les méthodes de Super-Résolution (SR) permettent de générer des images haute résolution (HR) à partir d'images basse résolution (BR). Contrairement à un rééchantillonage par interpolation, elles restituent les hautes fréquences spatiales et compensent les artéfacts (flou, crénelures). Parmi elles, les méthodes d'apprentissage automatique telles que les réseaux de neurones artificiels permettent d'apprendre et de modéliser la relation entre les images BR et HR à partir d'exemples.

Ce travail démontre l'intérêt des méthodes de SR à base de réseaux de neurones pour les systèmes de reconnaissance automatique. Les réseaux de neurones à convolutions sont particulièrement adaptés puisqu'ils peuvent être entraînés à extraire des caractéristiques non-linéaires bidimensionnelles pertinentes tout en apprenant la correspondance entre les espaces BR et HR.

Sur des images de type documents, la méthode proposée permet d'améliorer la précision en reconnaissance de caractère de +7.85 points par rapport à une simple interpolation. La création d'une base d'images annotée et l'organisation d'une compétition internationale (ICDAR2015) ont souligné l'intérêt et la pertinence de telles approches. Pour les images de visages, les caractéristiques faciales sont cruciales pour la reconnaissance automatique. Une méthode en deux étapes est proposée dans laquelle la qualité de l'image est d'abord globalement améliorée, pour ensuite se focaliser sur les caractéristiques essentielles grâce à des modèles spécifiques. Les performances d'un système de vérification faciale se trouvent améliorées de +6.91 à +8.15 points.

Enfin, pour le traitement d'images BR en conditions réelles, l'utilisation de réseaux de neurones profonds permet d'absorber la variabilité des noyaux de flous caractérisant l'image BR, et produire des images HR ayant des statistiques naturelles sans connaissance du modèle d'observation exact.

Remerciements

Je souhaite remercier toutes les personnes qui m'ont permis d'effectuer cette thèse dans les meilleures conditions. Merci à Christophe, Franck et Moez d'avoir été d'excellents encadrants et d'avoir su m'apporter des connaissance nouvelles et un support constant, toujours dans la bonne humeur. Je remercie également tous les collègues d'Orange pour leur investissement, leur conseils, leur aide et l'atmosphère de travail agréable. Je remercie l'INSA de Lyon et le laboratoire du LIRIS pour le travail administratif et d'offrir un cadre d'excellence pour l'accomplissement des travaux de thèse.

Sur un plan plus personnel je tiens à remercier les amis d'hier et d'aujourd'hui pour leur soutien. Merci aux copains de toujours, notamment à Thim et Johan qui ne manqueront jamais de ma rappeler que je suis ingénieur informaticien. Merci aux Zikets qui m'ont accompagné durant ces 8 années d'INSA et avec qui nous avons partagé bien plus que la sueur des veilles de partiels. Merci aux Rennais, fussent-ils colocs, musiciens, danseurs, Xpotes, collègues... j'ai passé 4 années enrichissantes grâce à chaque personne rencontrée sur ma route bretonne. Thanks to all the brummies that welcomed us with so much love, in particular the Slaters for their cosy house and lively family atmosphere !

Merci à ma famille d'avoir largement attisé ma curiosité scientifique et de m'avoir fourni un affectueux cocon pour la développer. Quel bonheur de voir surgir et grandir de nouveaux cocons ...

Enfin, merci Nadège d'avoir été un si encourageante et patiente, pour ce beau 8 avril et toutes les aventures que l'on n'a pas encore vécues.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

1.1 Context and motivations

The last twenty years have witnessed a spectacular advance in the democratisation of high technologies. Processes or devices that were expensive and usable only by specialists are now available for all of us and are part of our everyday life. In particular, photography and video acquisition devices are nowadays inexpensive and suitable for novices. The produced content is no longer stored on analogue supports but digitalized and sometimes synchronized with remote servers. Ericsson mobility report [Eri17] shows that 50% of the 8.8 ExaBytes of mobile data traffic in 2016 is due to video, and predicts that this ratio could represent 75% of 71 ExaBytes in 2022.

Retrieving information in such a huge mass of data requires automated process. That is why automatic indexation systems have been created and can address several scales, from personal to internet scale contents. Search engines can nowadays give instant results beyond keyword search, by a semantic understanding of the users requests. However, if a content is not manually annotated, *i.e.* when keywords or descriptors are not provided or irrelevant, classifying it is less straightforward. For multimedia content (images, videos, audio), this requires to have algorithms able not only to decode the given content in order to render it, but to actually recognise the content and provide annotation in order

to index it. Such processes may take place offline (*i.e.* on a pre-existing content) or online for real-time analysis systems since anybody can now broadcast live streaming on social media. Companies offering such services cannot afford to have human supervision given the massive scales (for example, Facebook reports 1.3 billion daily users), and struggle with inappropriate or illegal contents being streamed on their platform.

Usually, the general pipeline of an automatic recognition system can be split into two parts. For the first one, low-level signal features such as objects contours, edges or corners are detected, extracted and combined into intermediate representations that can describe object parts (*e.g.* a wheel, an eye, the stroke of a letter). The second part aims to discriminate the underlying representations to perform classification into several classes. To handle the different aspects of this pipeline, researchers have developed many solutions, involving signal processing, biologically inspired techniques, mathematical and statistical tools. On top of that, recent advances in machine learning and artificial intelligence have allow major breakthrough. Algorithms such as Artificial Neural Networks can be trained with large amounts of data and automatically learn how to extract relevant features and combine them to classify an object. The performance of such algorithms can surpass human on a number of tasks, as the machines can have access to millions of training samples.

Two important recognition systems for multimedia content indexation are Optical Character Recognition (OCR) and face recognition engines. The first allows to extract useful information such as speaker's name, locations, and video topic in conventional new reports for example. Many videos from social networks have now embedded subtitles to enable watching without sound which is a common usage of mobile devices. Face recognition can be used to recognise personalities in contents that do not contain textual information or annotation, or as biometric identification systems or create enriched photo collections.

At Orange, machine learning approaches and in particular Neural Networks for multimedia content understanding have been used since the early 2000's, applied to face detection and recognition [GD04, DG07], action recognition [BMW+11] or text detection and recognition [DG08, EGMS14]. In [Ela13], it was suggested that SR methods could be a potential technology to improve character recognition performance, as low-resolution is one of the bottleneck for OCR.

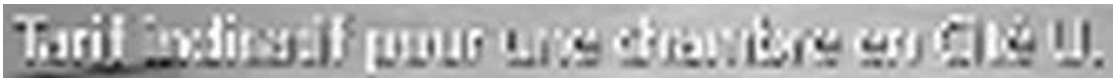
Several factors can deteriorate the performance of such recognition engines, either by ambiguous content or degraded image quality. Low resolution is one of the main operational challenging problems. It is caused by several reasons: distant objects, limited device capacities, digital image size reduction. In such case, the reliable features that

both humans and machines exploit to recognise shapes and objects end up being irrelevant. Fine characteristics such as textures, are mixed into single pixels. Other specific features are lost, such as those allowing to discriminate between two individuals in face recognition or fine spacing differences between "nn" and "m" in character recognition. Such examples of low-resolution content are displayed in Figure 1.1. As a consequence, both human perception and machine vision systems are flawed because unprepared for such badly shaped images.

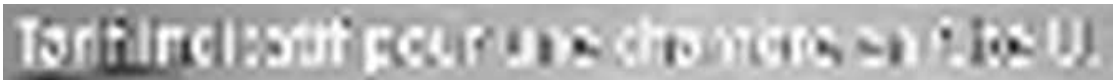
(A) Original HR image

(B) LR image, $s = 2$

(C) LR image, $s = 3$



(D) LR image, $s = 4$



(E) LR image, $s = 5$

FIGURE 1.1: Text image extracted from a TV stream, synthesised at several resolutions using a downsampling factor s. The LR images are upscaled using bicubic interpolation to have the same size as the original one.

Inspiring researchers – and also fictions, see Figure 1.2 – methods designed to recover a High Resolution (HR) image from one or several Low-Resolution (LR) are known as image Super-Resolution (SR). This topic has been very active since the publication of Tsai and Huang [Hua84] who first modelled the problem. These solutions can produce SR images that are much more pleasant than a simple interpolated one: they compensate jaggy edges or blur artefacts by modelling the relationship between LR and HR images. Several strategies have been adopted to achieve such improvement. Multiple Image SR techniques assume that many LR images of the same scene or object are available. As each image contain a slightly different version of the real scene, it is possible to merge this scattered visual information into one, enhanced image. When only one LR observation is available, the problem is reduced to a highly ill-posed inverse problem called Single Image SR (SISR). As the information is not scattered in several images, these approaches use external knowledge to improve the resolution. This thesis will focus on this second category of methods, using machine learning algorithms to bring this external knowledge into the proposed SR frameworks

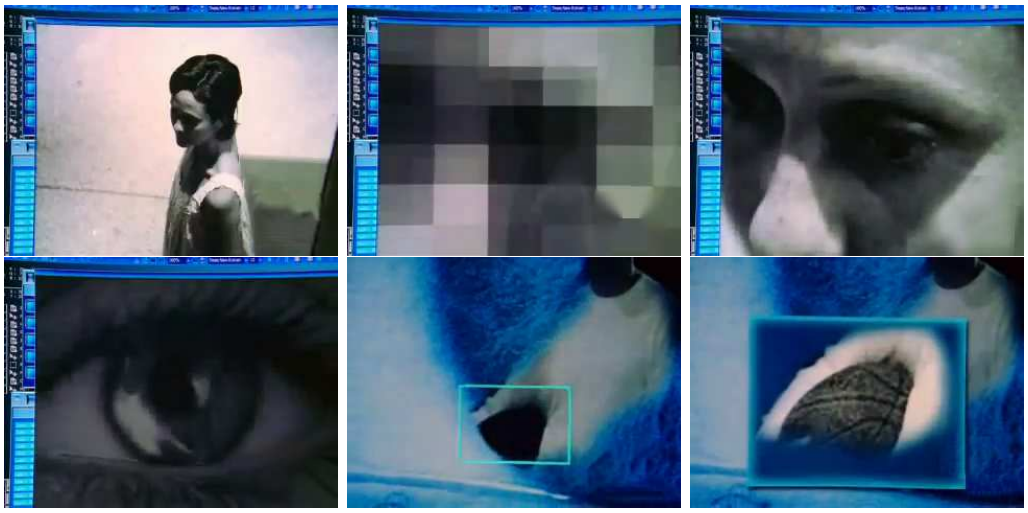


FIGURE 1.2: An illustration of SR in popular culture, from the CSI serie[1]. While the investigator states "*Magnification one hundred for starters*", a piece of forensic software allows to zoom on the eyes of a witness to reveal a critical clue.

SR has become a standard subject to illustrate advances in machine learning and artificial intelligence, with a communication around this subject that goes far beyond the scientific community, as anyone using social networks or portable devices have come across to visual contents with limited resolution.

Machine Learning has also brought new ways to perform SR. Since the influential work of Baker et al. [BK02] in 2002, example-based approaches for SR are the most competitive methods. They aim to capture the relationship between LR and HR image spaces, by algorithms that learn from many examples of low and high resolution image sets. For instance, algorithms such as neural networks, dictionary learning, manifold learning or sparse coding have been used (see chapter 3).

1.2 Scope and objective

The purpose of this thesis is to propose example-based Single Image SR (SISR) approaches to improve both image quality and the performance of recognition systems. The SR images must therefore contain the same relevant features as HR images, matching the prior of the recognition system. While other approaches seek to design recognition systems able to directly recognise low-resolution content, developing SR approaches has several advantages:

1. It can be easily integrated in a existing production framework and takes the form of an independent technological component.

[1] http://www.imdb.com/title/tt0247082/

2. Second, the same component may be deployed on different systems with similar needs for high-resolution images, without redesigning each individual recognition system.

3. Third, it allows to have an explicit representation of the intermediate result, and use at least two criteria to evaluate the benefits: how well the image is reconstructed, and how beneficial it is for the downstream system.

In particular, we will focus on improving OCR and face recognition performance using SR. To do so, we will address text and facial images SR, with different approaches.

This work also aims to address SR in realistic contexts. Usually, SR methods consider that the LR images are obtained from a HR image with a fixed observation model. However, LR images extracted from multimedia content may have undergone more complex degradation. While approaches based on the observation model can account for such variations, it is more difficult with example-based approaches that learn from the data itself, often via a implicit mapping. If the examples does not reflect the variations of the real world, or is the algorithm cannot handle the diversity it faces, no suitable approach can be trained.

1.3 Summary of our contributions

This thesis is divided into three main contributions. First, to address text image SR in several contexts (document image, texts from televisual contents), a method based on neural networks is proposed. This example-based approach relies on specific datasets that provide useful samples of LR and HR images, and deep neural architectures that allow to efficiently capture the relationship between the LR and HR image spaces. It is designed to improve both the image resolution and the performance of automatic recognition systems (OCR). A new dataset for single text image SR is presented, from which the first international competition on super-resolution was organised. Results of this competition are reported and analysed.

The second contribution consists in a new approach for face SR, in order to improve facial recognition engines. The method consists in two steps performed by neural networks: a first generic step that improves the resolution of the whole LR image followed by a specific step that focuses on the facial components such as the eyes, the nose or the mouth. This hierarchical approach allows to incorporate an *a priori* knowledge of the automatic face recognition systems – the dependency on high-resolution facial features.

The third contribution is related to the following real-world observation: while example-based approaches are very efficient for SR on synthetic low-resolution images (*i.e.* obtained from a high-resolution image with a known observation model involving blur and downsampling), it is not likely to process images for which this observation model holds. To handle the diversity of LR images that can be found in real-world application of SR, a deep convolutional neural network is trained on a large database created obtained with many different observation models.

1.4 Organisation of the manuscript

The rest of this manuscript is organised as follows:

- Definitions and SR application domains are presented in chapter 2.
- A literature review is presented in chapter 3. A particular attention is paid to example-based methods and domain-specific approaches for text and face images.
- In chapter 4, the first contribution, based on neural networks, is introduced for text images from documents and TV streams. Results of the first international Text Image Super-Resolution competition of ICDAR 2015 and the associated dataset are also reported.
- Leveraging on the first method, the second contribution described in chapter 5 is new approach to facial SR. It associates a local generic model followed by a face-specific step, using two neural architectures.
- The third contribution in chapter 6 addresses the variability of realistic imaging models. Using a Deep Convolutional Neural Network, results demonstrate that a robust model can be trained to perform SR on a variety of low-resolution observation models.
- Finally, a summary of the contributions and perspectives are presented in chapter 7.

Part I

Definitions and Literature Review

Chapter 2

Definitions and Application Domains

Contents

2.1 Introduction

In this chapter, we first give relevant definitions in section 2.2 that will constitute our basis for the following of the report. We start with defining the terms commonly used in SR, then describe the use of interpolation methods and their limits, and finally describe the SR imaging model used in the literature and throughout this work.

The different application domains are reported in section 2.3, including visual enhancement, security and preprocessing images for improving recognition systems, which is the main application addressed in this thesis.

2.2 Definitions

2.2.1 Defining resolution: spatial and frequency aspects

The term of resolution is a first struggle when addressing the SR domain. While distinct, terms such as resolution, definition, quality or density are often mixed together. This is comprehensible as they describe very close properties of signals, actuators or sensors - in particular images. However, they do have a precise signification that we will point out and illustrate in the following.

- The *resolution* of a discrete signal is related to its sampling frequency. Therefore, a high-resolution image is an image with a high pixel density. It makes more sense when considering scanning devices, where calibration is more precise: the resolution can be defined in pixel per inch (*ppi*) which directly links the density to an absolute quantity. The resolution power of a sensor refers to its capacity to distinguish between two ideal point sources.
- The image *definition* refers to the number of pixels of an image. However, it is often related to the image quality, as resizing an image is a very common practice for display adaptation for example. Thus, a high definition video is also a video that was recorded using a camera able to hold a good enough resolution to produce an image content in coherence with such a frame size.
- More broadly, *image quality* is a term that covers a large spectrum of concepts. Generally speaking, it is a perceptual notion based on the Human Visual System (HVS) that can be measured on several scales, using different procedures like sharpness or object recognizability. It may also be evaluated automatically, with or without a reference image, depending on the employed measure. It also covers the field of video processing, taking into account the temporal receptivity of the human brain, that can be more important than the per-frame quality in some cases.

Image Super-Resolution therefore relates to methods that allow to increase the *resolution*, *i.e.* to resample an image at a higher sampling rate. The "Super" prefix indicates that these methods must also provide an accurate image signal that is not subject to distortions. To simply augment the pixel density without concern of the resolution (*e.g.* the distinction of two points spots) of the content, most common practice is to use interpolation. Thus, before addressing SR, which also aims to recover a latent high-resolution image, we review the common interpolation approaches in image processing.

2.2.2 Resampling and Interpolation in Digital Image Processing

Interpolation methods allow to reconstruct a continuous signal from a series of discrete points. In still image processing, these points are generally uniformly sampled on a 2D grid. However, depending on the acquisition method and the treated data, one can be led to process non-uniformly sampled signals. For instance, the fusion of multiple aligned images can be realised by interpolating the aligned pixels that constitute an irregular grid, followed by a uniform resampling on the desired regular one.

Such a reconstructed continuous signal can then be sampled at a different rate. If this rate is higher than the original signal rate, one is performing discrete interpolation or upsampling ; if the rate is lower, one is performing decimation, subsampling or downsampling. Those different terms can have some preferential usage depending on the context and the used technique.

Interpolation for fixed sampling rates can be seen as a convolution of a continuous kernel with the discrete samples (see Figure 2.1). The kernel can be seen as the impulse response of the chosen interpolating method. We will not consider the causality of this impulse response, as it has varying interest depending on the considered kernel. Additionally, finite impulse response are practically used.

Once a continuous signal is reconstructed, it can be sampled again to upscale or downscale the image signal. This is illustrated in Figure 2.1 via the multiplication of a continuous signal with a Dirac comb, that yields a discretised signal at a different sampling rate.

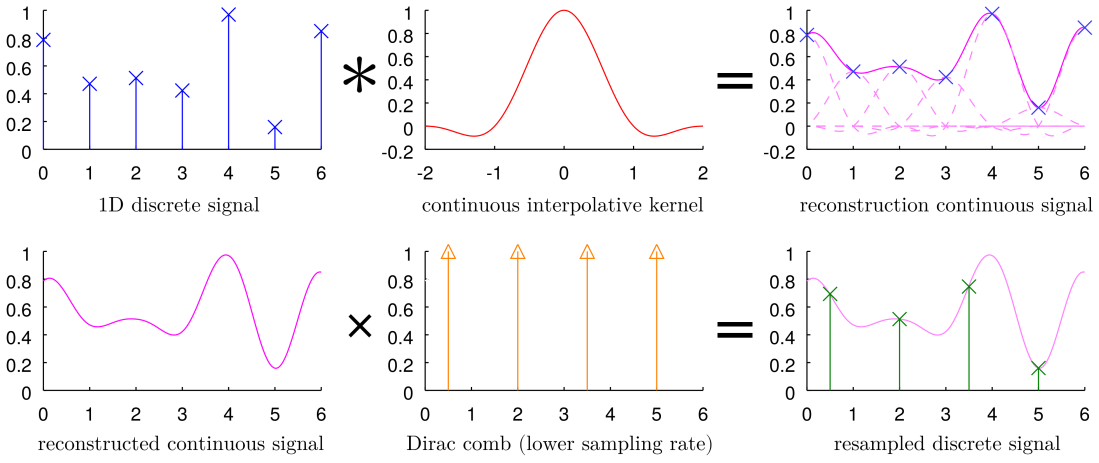


FIGURE 2.1: Resampling via continuous reconstruction of a discrete 1D signal. First line: the input signal (blue) is first transformed into a continuous one (pink) using an interpolative kernel (red, here Lanczos-2). Second line: the resulting signal (pink) can be resampled at another sampling rate (here, $^2/_3$ with an offset of 0.5 to illustrate a case of resampling with non-integer ratio). This operation is modelled by a multiplication of the continuous signal with a Dirac comb (orange).

In most implementations of resampling methods, the continuous signal is not reconstructed. Instead, discrete convolutions or matrix multiplications are used. The most used interpolation methods in image processing and Super-Resolution are generally referred to as nearest neighbours, linear, cubic, Lanczos($-n$), Gaussian. Several optimisations (sample-wise computation, discrete convolutions, pixel recopy) ensure efficient implementations of those algorithms for the overall interpolation procedure. In the following, we describe each method, and then sum them up in table 2.4.

2.2.2.1 Reconstruction of ideal unaliased signals with sinc interpolation

Using *sinc* reconstruction aims to reconstruct a "perfect" continuous signal, from a discrete signal obtained under Shannon-Nyquist aliasing conditions [Sha49] stating that the original spatial sampling frequency ω_s is at least two times higher than the maximum spatial signal frequency ω_{max}:

$$\omega_s \geq 2 \times \omega_{max} \tag{2.1}$$

For a given discretised signal $i(x)$ (*i.e.* a continuous signal multiplied by a Dirac comb), its Discrete Time Fourier Transform (DTFT) spectrum will have an infinite repetitive nature, every $2 \times \omega_s$ as depicted in Figure 2.2, as a Dirac comb in the spatial domain is also a Dirac comb in the frequency domain.

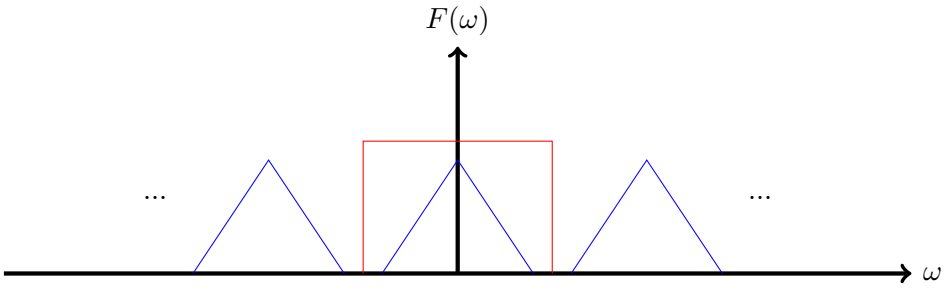


FIGURE 2.2: The DTFT of a discrete 1D signal (in blue) is composed of a repetition of the continuous signal spectrum every ω_s (the sampling frequency). To recover the FT, the DTFT must be multiplied by an ideal low-pass filter (in blue) to retain only the central spectrum.

To get back the continuous signal, we need the original spectrum by cutting out the DTFT one at $\omega_s/2$. This means multiplying the spectrum by a *rect* function, centred in $\omega = 0$ and of width ω_s. The inverse Fourier transform of such *rect* function being a normalised *sinc* and the equivalent of a multiplication in Fourier domain being a convolution, we might obtain a continuous signal $I_c(x)$ by its convolution with our

discrete signal:

$$\begin{aligned} I_c(x) &= i(x) * sinc\,(x) \\ &= \sum_{n\in\mathbb{Z}} i(n) \times sinc\,(x-n) \end{aligned} \tag{2.2}$$

which is a sum of scaled *sinc* functions. Although *sinc* interpolation can theoretically reconstruct a perfect signal, it is limited by at least two aspects: it has a non-causal infinite impulse response, and images – in particular LR images for SR – may not be compliant with the Nyquist requirement and therefore contain levels of aliasing. Thus, interpolative kernels that exhibit a similar behaviour as the *sinc* function are used for practical interpolation.

2.2.2.2 Practical interpolation methods

In practice, intuitive and more simple interpolation schemes are used, predicting missing pixel values from its close neighbourhood. We adopt a convolution point of view; although block circulant matrices or loops can be used for actual implementation.

- **Nearest neighbour** interpolation resamples an image by assigning the value of the nearest sample to the processed samples of the new grid. This process can be seen as a continuous reconstruction by convolution of a *square* function with the discrete signal.
- **Bilinear** interpolation computes the value of a pixel on the new sampling grid by a weighted sum of the 4 nearest neighbours. The weights are computed according to the euclidean ($L2$) distance. Alternatively, this can be seen in $1D$ as a convolution with a symmetrical *triangle* function (see Figure 2.4).
- **Bicubic** interpolation is a widely used method (for both upsampling and downsampling) that defines a convolution kernel $p_a\,(x)$ with the following polynomial formula (for one dimension x):

$$p_a\,(x) = \begin{cases} (a+2)|x|^3 - (a+3)|x|^2 + 1 & \text{for } |x| < 1 \\ a|x|^3 - 5a|x|^2 + 8a|x| - 4a & \text{for } 1 < |x| < 2 \\ 0 & \text{elsewhere} \end{cases} \tag{2.3}$$

  It is implemented in most of the image processing libraries, in particular in Matlab with $a = -0.5$ and the OpenCV library with $a = -0.75$. The $2D$ kernel can be seen in Figure 2.4.

- **Lanczos−n** interpolation (n being the neighbourhood in pixels) is based on the following formula for the kernel $p_n(x)$ (for one dimension x):

$$p_n(x) = \begin{cases} sinc(x) \times sinc\left(\frac{x}{n}\right) & \text{for } |x| < n \\ 0 & \text{elsewhere} \end{cases} \tag{2.4}$$

  where *sinc* is the *sinus cardinal* function. It aims to reconstruct realistic signals as it uses *sinc* function (perfect reconstruction under Nyquist requirements, see previous paragraph), but modulates it by a second low-frequency *sinc*. The $2D$ kernel can be seen in Figure 2.4.

Figure 2.3 depicts the ($1D$) spatial and frequency aspects of the mentioned kernels. We can note that similar spatial kernels can exhibit noticeable differences in frequency. The fig ures are obtained from time-limited signals, which explains the non-*rect sinc* frequency spectrum.

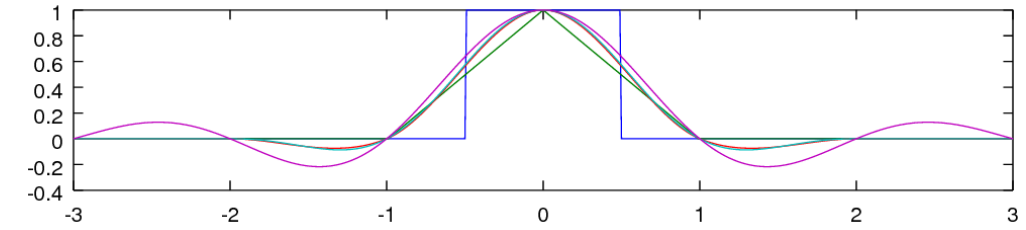
2.2.2.3 Digital processing and Kernel Sampling

We have shown the continuous functions used for resampling and presented the theoretical two-step approach. Practically, the kernels are usually sampled, and the desired image is obtained by convoluting the sampled kernel with the original images. However, we are left with a choice for the sampling phase, *i.e.* whether the sampling occurs on or off the original sampling grid. Figure 2.5 illustrates the different choices of phase.

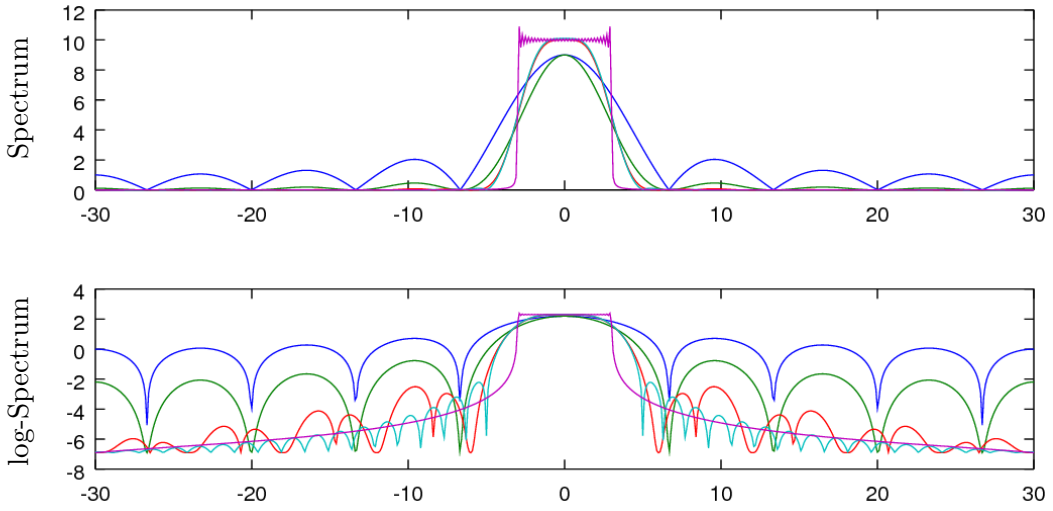
One important thing to remember is that the choice of resampling grid or the phase is crucial in some evaluation scheme, as it can produce subpixel shifts and misalignment. Indeed, the common measures for evaluating depend on Figure 2.6 illustrate the consequence of the chosen phase.

2.2.2.4 A note on subsampling

Generally, when resampling at a lower spatial frequency to simulate a low-resolution observation, a low-pass filter is applied along the convolution kernel to prevent an strong aliasing effect to take place. However, if no aliasing occurred, the loss of spatial resolution would not play any role and the problem would be equivalent to a deblurring problem. A discussion on this aspect will be conducted later in this manuscript in chapter 6. The reader can also refer to [Tur90] for advanced frequency analysis of the aforementioned filters.



(A) Spatial Comparison of the different interpolative kernels



(B) Corresponding Spectra and log-Spectra of the selected kernels (on a different bin scale for a better behaviour visualization).

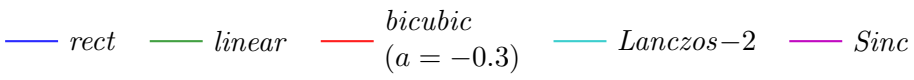


FIGURE 2.3: Various $1D$ interpolative kernels in the spatial (A) and frequency (B) domains. The kernels that mimic a *sinc* kernel such as *Lanczos*-2 (cyan) or **bicubic** (red) are more likely to approach the ideal low-pass filter approached by the *sinc* (purple).

2.2.3 Image observation model for SR

Image observation models are used to represent the process between a source scene and a final discrete image. This process may incorporate the different factors that influence the transmission of the light signal in the different channels (analogue and digital). We can end up with interaction at a physical level, in the scene (atmospheric or illumination conditions), or inside the device (lens aberrations, sensor defects, electromagnetic noise). After discretisation, some factors can still interact with the production of our final image (quantization, bit errors, compression). Not all those factors are taken into account, especially for SR. In the following, we shall first consider continuous aspects (how an image is acquired from the continuous scene) and refer to works that rely on this aspect.

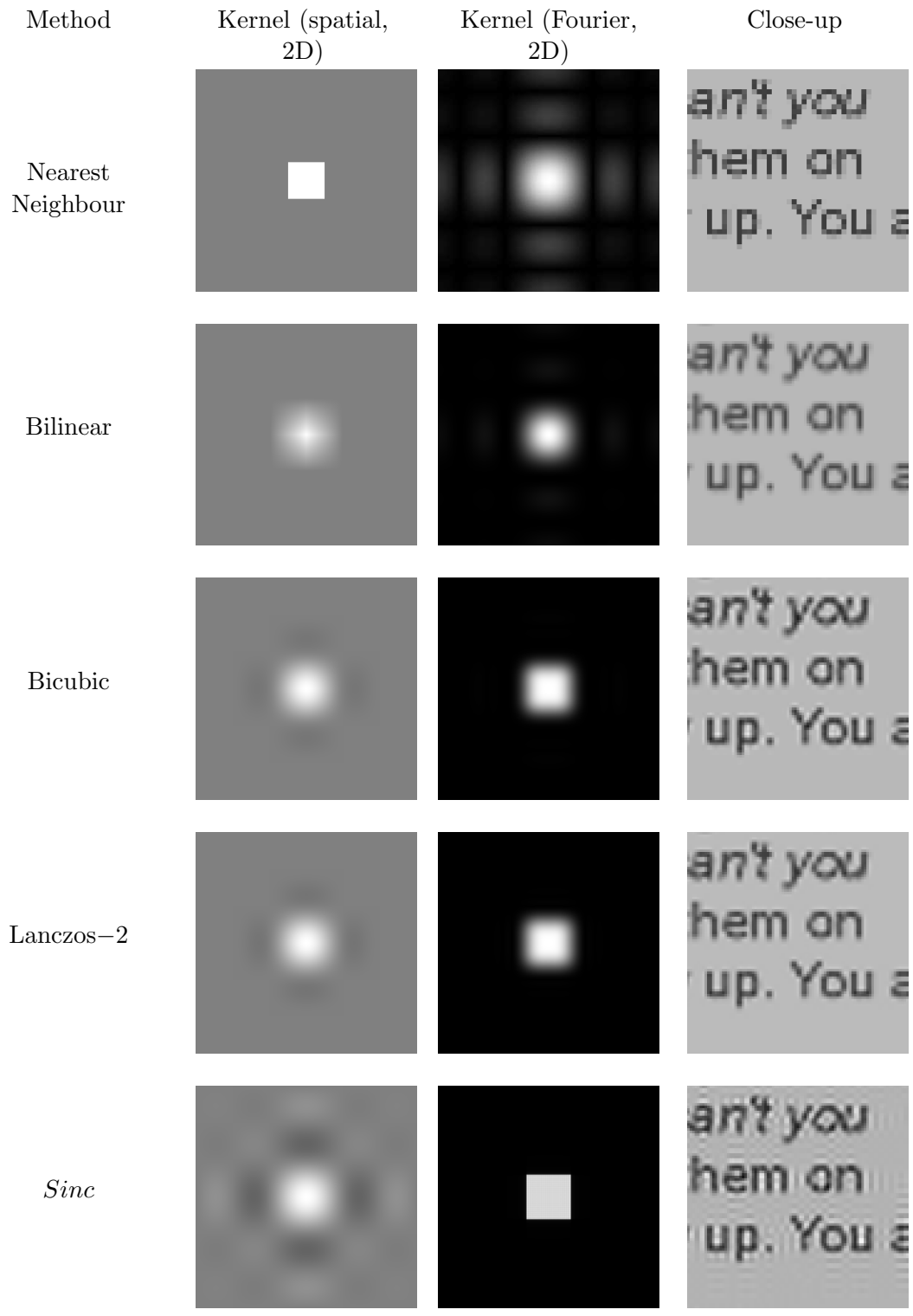
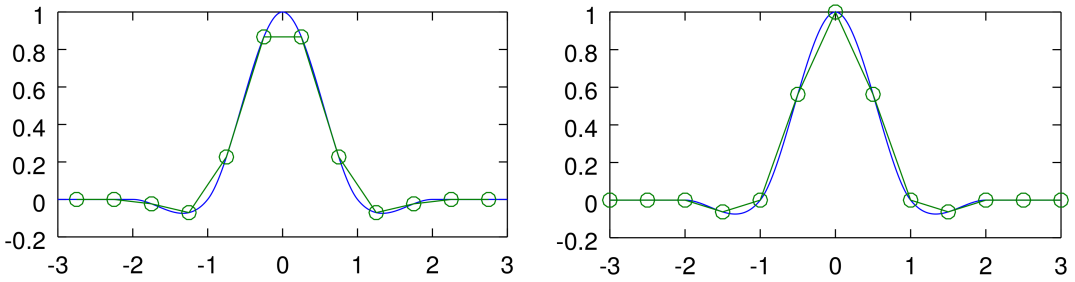


FIGURE 2.4: Interpolation kernels with their spatial and Fourier appearance, and an interpolated image sample, for a scale factor of two. Artefacts are clearly visible: nearest neighbour interpolation produces blocky images as the pixels are simply reproduced, bilinear interpolation creates oversmooth images. Bicubic and Lanczos interpolations are less smooth, but tend to produce overshoot and ringing artefact on strong edges. As the low-resolution image contains strong aliasing and strong edges, the *Sinc* reconstruction results in many ringing artefacts.



(A) Off-grid kernel for downsampling by even factors

(B) On-grid kernel for downsampling by even factors

FIGURE 2.5: Different choices of 1D interpolative kernel sampling for digital downsampling, illustrated here for a downsampling factor of 2 and a bicubic kernel. The blue curve is the continuous kernel, the green one is the obtained discrete version. With the first kernel, two samples will be merged with an equal value while with the second, every other pixel will less contribute to the downsampled signal.

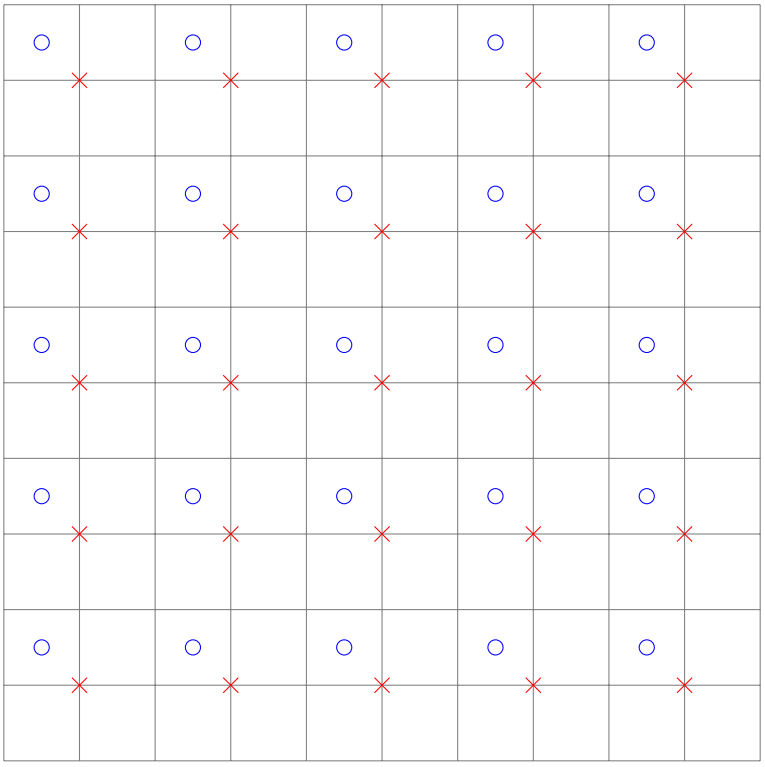


FIGURE 2.6: Different choices of downsampling grids compared with the original HR pixel grid (on-sample for blue circles, off sample red crosses) that lead to the same resolution (number of pixels in the image) but with a different subpixel alignment.

Then, we shall focus on discrete-to-discrete models for the creation of low-resolution images.

2.2.3.1 Continuous observation models

A perfect SR method should be able to recover the original continuous scene. However, this is impractical for several reasons: the infinite possible scenes that produced the observation (even with the use of regularization or constraints) and the non-storable nature of such continuous signal. However, the continuous parameters such as the point spread function or the continuous latent image can be manipulated to model the problem. Although SR methods often consider a discrete inverse problem representation, some works such as [MI13] consider a real-world set-up and the influence of the choice of different Point Spread Functions in the model.

2.2.3.2 Discrete observation models

Discrete models are used in most of the works of SR. They consider low-resolution image as being obtained from a high-resolution image via a pipeline displayed in Figure 2.7. Note that the definition of HR images are sometimes ambiguous. Generally, the HR images corpora are selected so that they do not exhibit artefacts (compression, blurry areas). However, ensuring that all the images have desirable HR content is harder in large scale image databases.

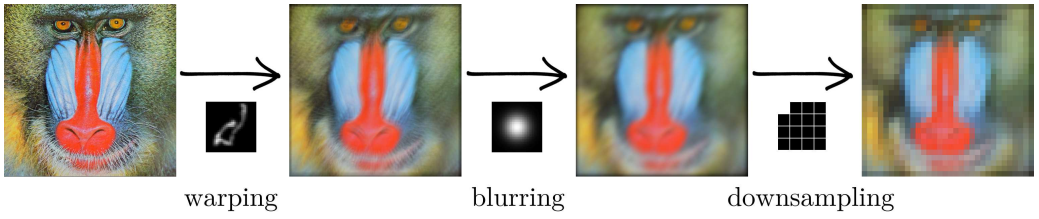


FIGURE 2.7: Discrete observation model for LR image synthesis from HR images.

Warping The warping process allows to model an inter-image variability that is important in the case of multiple image SR. For SISR, this step is skipped and we end up with a simplified process (see Figure 2.7). For warping, various strategies are adopted, inducing different level of complexity. The following enumerates the most common transformations.

- **Affine transformations** – the warped image is obtained through linear transformation of its coordinates.
  - shift: real or integer-valued shifts can be randomly introduced in the 2D HR space (or grid). This simulates real-life behaviour such a subject translation

or hand shake. The transformed image $\mathcal{I}$' is often expressed from the original image $\mathcal{I}'$ as follows:

$$\mathcal{I}' = \mathcal{I}\left(x + \Delta x, y + \Delta y\right) \tag{2.5}$$

where (x, y) are the 2D coordinates and $(\Delta x, \Delta y)$ are the relative motion. Some coarse to fine methods can be employed to reduce a large-scale motion or translation to a subpixel motion estimation problem. The transformation matrix in homogeneous coordinates is:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \Delta x \\ 0 & 1 & \Delta y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

– rotation: with the same objective, and assuming a small rotation angle $\Delta\theta$, a rotation can be added, so that

$$\mathcal{I}' = \mathcal{I}\left(x - sin\left(\Delta\theta\right), y + sin\left(\Delta\theta\right)\right) \tag{2.6}$$

The transformation matrix in homogeneous coordinates is:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} cos\left(\Delta\theta\right) & -sin\left(\Delta\theta\right) & 0 \\ sin\left(\Delta\theta\right) & cos\left(\Delta\theta\right) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \simeq \begin{bmatrix} 1 & -sin\left(\Delta\theta\right) & 0 \\ sin\left(\Delta\theta\right) & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

– General affine and perspective: a more general spatial transformation can be assumed, as in [CZ01], to simulate perspective that can occur in real-life sequences:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

where a_i and b_i are the coefficients of the affine transformation.

- **Other transformations**

  – non-rigid: in some applications, non-rigid motion can be taken into account, like in [YB08] for facial expression change.

– occlusions: while recovering a latent or hidden object is rather addressed by inpainting, the short or long-tem occlusions in time may be taken into account in the models by outlier regions detection (e.g. [FKMI09]).

However, some methods do not assume such parametric approaches or do not include them in their model, such as [PETM09] (see paragraph 3.3.3.3 of chapter 3 for more details). Instead, they perform a non-local search based on similarity between patches to approximate the alignment (see section 3.2 of chapter 3).

Blurring Blurring is an important part of the process as it removes high-frequencies and avoid to end up with a totally aliased low-resolution signal. In the SR model, it corresponds to the point spread function (2D impulse response) of an optical device, but may also account for the other physical phenomena such as atmospherical blur or defocus. However, this stage does not include other physical interaction and aberration such as coma or astigmatism. Specific works address these problems in the literature, but not explicitly via Super-Resolution.

Different strategies have been proposed for blurring. Most of the time, it is performed via a low-pass filter convolution kernel. To imitate natural non-negative single channel blurring kernel, the most employed kernels are tap (or box) kernel, Gaussian kernel of diverse mean and variance and round circle (out-of-focus blur). However, kernels with negative values are also used. Mainly, this case arises with the use of interpolative kernels such as the bicubic and Lanczos ones. Their implicit goal is to reconstruct a realistic continuous signal, *i.e.* approaching a *sinc* with a finite definition interval. As shown on Figure 2.4 in subsection 2.2.2, negative lobes exist. However, for downsampling (see next paragraph), these kernels may be anti-aliased *i.e.* have a 2 times larger spatial width (2 times smaller spectral width, thus LPF). Furthermore, in [MI13], authors reveal that positivity might not be a suitable constraint for the blurring kernel in SR, as they demonstrate that the optimal kernel found by a blind deconvolution process does present negative lobes.

Downsampling This last step, which can also be referred to as decimation or sub-sampling, is what makes SR fundamentally different from related linear problems such as deblurring, as it produces a spatial dimension shrinkage. Most of the time, this last step is simply a pure decimation process, which is taking every n other pixels. However, cases can arise where a non-integer scaling factor may be needed. In this case, interpolation methods with resampling (see subsection 2.2.2) are necessary and can not be modelled as a pure filtering and decimation operation.

Noise An additional noise term can be added. In many methods, only a low level of noise is considered as the main objective is to recover the lost information during the warping+blurring+downsampling process. However, a quantization noise is often present during actual image saving and loading operations. Some work also add gaussian noise.

2.3 Application Domains

Image SR approaches may serve several purposes, from image restoration to forensic analysis. Other signals may also benefit from higher sampling rates. The following paragraphs describe various applications of SR in different domains.

2.3.1 Image refinement and visual enhancement

Many works, especially those addressing general-purpose SR and natural images, can be used to synthesise HR images from low-resolution ones in order to provide better shaped images to consumers or users. Dedicated approaches give a particular attention to the computation load and speed, considering that such approaches might be applied to real time image resizing.

2.3.2 Surveillance and security applications

In spite of the high-definition sensors that are available and that equip more and more devices, the problem of low-resolution content still arises in many sensible contexts like in military or surveillance applications. This can be due to outdated devices, the distance of a subject to the camera, atmospheric or out-of-focus conditions, subject or camera movement and many other reasons. Super-resolution is a way of overcoming such problems and is very critical if used as court proof.

2.3.3 Pre-processing for automatic recognition system

SR can also be used as a pre-processing step, between a captured image and an automatic processing unit that has an *a priori* of high-resolution images. This can be the case of many recognition systems such as Optical Character Recognition (OCR) or Face Verification systems. In this case, while a good image reconstruction can still be interesting, the objective is to produce high-resolution images in which a recognition system can detect or recognise objects better.

This thesis mainly addresses this category of applications. SR approaches for several types of text images SR are presented in chapter 4, and an approach improving facial images resolution is proposed in chapter 5.

2.3.4 Other paradigms

Audio Super-Resolution Some works have been conducted on the super-resolution of audio signal. This is highly related to bandwidth extension with many applications since the late 90's in speech (notably speech over the phone), music and audio in general [DBBE+09]. More recently, inspired from the strong interest in image SR, some authors have proposed several methods to address audio SR. As in image SR, the interest is to recover high-frequency lost during acquisition or because of the transmission channel constraints. In recent years, sampling rates for audio have also increased in commercial products, with the emergence of labels such as Blu-ray Pure Audio or Hi-res audio, that propose sampling rates from $96KHz$ up to $192KHz$, with higher quantization too (24 bits, where traditional quality is $44.1Khz$/16 bit). Analogue to the SD to HD conversion for video, audio SR can be a solution for conversion of audio signal at such sampling rates.

Another application in audio is spectrogram Super-Resolution. The traditional way to obtain spectrograms is to apply Short Term Fourier Transforms (STFT) on successive windowed parts of the audio signal. However, due to the Heisenberg singularity, increasing temporal resolution (*i.e.* smaller windows) will decrease the frequency resolution, and vice-versa. Super-Resolution, in this context, can offer a way to overcome this limitation and increase the resolution of one of the axis, as in [NMG+10].

Gray-scale Super-Resolution (re-quantization) In [HSCG10], the authors study the impact of "low grayscale resolution', *i.e.* images of poor contrast or low quantization scale. This case may arise in poor lighting condition as for contre-jour shooting. Finding the intermediate grayscale values can be seen as super-resolution of the quantization, and is closely related to the works on High Dynamic Range (HDR) imaging.

Super-Resolution in Microscopy Super-Resolution microscopy is a set of techniques that allow to increase the resolution of images captured by microscopes, going beyond the diffraction limit. Even if not applicable to every kind of image and devices because dealing with specific images and conditions, these techniques effectively increase the resolution of images by exploiting special behaviour of light emitters in biology microscopy. Those method can be classified into two groups:

- Deterministic super-resolution: exploits the non-linearity of fluorophores responses to excitation to resolve emitters.
- Stochastic super-resolution: exploits the non-stationary temporal behaviour to resolve emitters.

"Semantic" Super-Resolution An interesting concept is proposed in [NS08], where a semantic description of an event (*e.g.* a news event) can be "super-resolved" by gathering different sources of information, particularly in the case of audiovisual content in the patent. Aggregating information from different sources in order to provide exhaustive information can be an interesting research field, with challenging problems such as concordance, source trust, data aggregation.

2.4 Conclusion

This chapter recalls basic definitions that are useful in the context addressed by this thesis. Interpolation methods allow to resample a digital signal to different sampling rate. In the case of an image, they can be used to decrease or increase its resolution. However, when the resolution is increased by interpolation, the resulting image intensity values are a fixed linear function of those of the low-resolution image. Therefore, only the low spatial frequencies are reconstructed in the new image which lacks of sharp details and other high frequency contents.

SR techniques aim to go beyond these simple resampling methods and predict the missing high frequencies. For SR, the LR images are considered to be obtained from a HR image undergoing successive transformations: warping, blurring and downsampling.

Such techniques may be used for many image processing applications, including natural image enhancement, surveillance and pre-processing so that automatic recognition systems see their performance improved on low-resolution content. This last point will be the core of several contributions presented in this manuscript.

Chapter 3

Literature Review

Contents

3.1 Introduction

Image Super-Resolution is a research topic that has drown a lot of attention since the term initially appeared in the early eighties. It refers to a large set of methods that aim to reconstruct high resolution images from low resolution ones. Not only those techniques increase the spatial pixel density, but also recover missing high-resolution information, either by gathering it from several observations or by inferring it from a structured *a priori* knowledge. Depending on the processed data and the problem modelling, several techniques have been explored through the years. The first historical methods addressed the Multiple-Image Super-Resolution problem (MISR) where several observations of the

same scene are available, with slight variations. Under certain conditions, the high-resolution image can be recovered almost perfectly. In this survey, we will present the main methods and refer to more advanced works on this subject. For Single Image Super-Resolution (SISR), recovering a high-resolution image necessarily involve an external knowledge, as HR information could not have been split into several observations. The available information is therefore very poor. This is why the SISR methods aim at establishing a relation between LR and HR images spaces.

This literature review starts with an overview of MISR methods, which is not addressed in this work but still very important in the SR community. Then, we propose a more extended review of existing methods in single image SR, with a focus on learning-based (or example-based) methods, which is the category of method our contributions rely on. Finally, we focus on specific SR methods which are designed and evaluated in order to serve other image processing tasks such as object recognition, as this is how we orientated our work.

This chapter is organised as follows. A short review of Multiple Image SR methods is provided in section 3.2. A more advanced review of the Single Image SR literature can be found in section 3.3. In section 3.4, a focus is made on approaches that are specific to text and facial images, as they are the category of methods that we address in this work. Finally, the scope of this thesis is defined regarding the literature review in 3.5.

3.2 Multiple image SR

Super-Resolution was first proposed as a set of methods for synthesising a high-resolution image from multiple (K_l) low-resolution observations, referred to as Multiple-Image Super-Resolution (MISR) in the following. These observations are supposed to contain a slightly different version of the same scene due to perspective, rotation, translation, scale and noise. The term Super-Resolution is re-employed for the singular case of $K_l = 1$, referred to as Single Image Super-Resolution (SISR). As stated in the introduction of this chapter, the SR problem can be seen as a linear inverse problem where a high-resolution image x (or more conceptually, a scene) has been transformed through a linear transformation T into K_l low-resolution images y_i. These linear transformations include spatial dimension reduction, which yields low-resolution observation vectors of size $M^2 < N^2$. For this inverse problem to be well-posed, we need a number K_l of low-resolution observations, sufficiently large so that $K_l \times M^2 \geq N^2$. Still, the problem can be ill-conditioned (not stable under disturbed input) and require additional constraints to render a final, artefact-free HR image.

In this thesis, we address single image super-resolution. However we give here a short review of the MISR methods as SISR can sometimes be performed using the same methods, with $K_l = 1$, K_l being the number of LR observations.

3.2.1 General principles

In general, MISR is addressed as an inverse problem, where LR observation have been obtained following the procedure depicted in Figure 3.1.

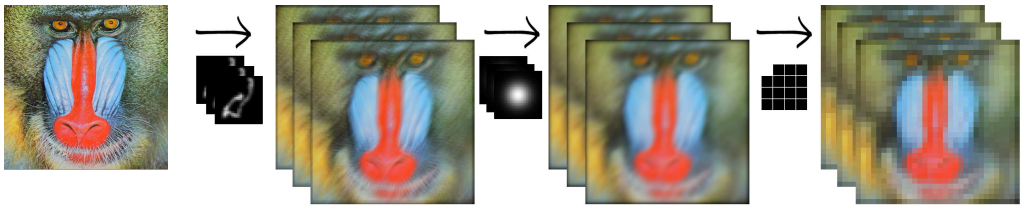


FIGURE 3.1: Observation model for multiple low-resolution images obtained from the same high-resolution image.

Each image is supposed to be obtained from the HR scene that must be reconstructed. The aim is to gather the split information in the different observations. For this, the most frequent approaches reverse the imaging model by finding the relative motion of the LR frames, fuse them and remove remaining artefacts such as blur. These steps can be done separately or jointly in an iterative scheme.

3.2.2 Main Approaches in Multiple Images SR

The following gives an overview of the different approaches to perform the steps presented in the previous paragraph.

3.2.2.1 Image registration

Most of MISR methods highly depend on registration techniques. This registration step is crucial to allow the fusion of the different sources of split information. As mentioned in subsection 2.2.3, different motions are taken into account. Image registration is a broad research subject that impacts many application fields such as compression, medical imaging or remote sensing.

Global image registration As mentioned in the observation model paragraph (see 2.2.3.2), a coarse-to-fine approach may be considered to first account for large translation, and then subpixel alignment matters. In [GO04], the authors employ area block matching to select the area that are most likely to be registered.

Local subpixel registration For subpixel registration, several strategies can be used. However, the most common is to interpolate all the observed images and find the minimum of a pixel-wise error mesure (such as the sum of squared differences):

$$e_{reg} = argmin \parallel I_{ref} - f(I_k) \parallel \tag{3.1}$$

After the registration stage, the aligned information has to be fused into a single image.

3.2.2.2 Image fusion

Once images have been aligned, they must be somehow fused to aggregate the scattered information. Basic methods referred in the survey [NM14] include mean or median filtering, weighted mean filtering, iterative back-projection, SVD-based fusion, pixel-wise Adaboost classifier, projection onto convex sets, Maximum Likelihood and Maximum A Posteriori (Bayesian) methods. In the next paragraph, we focus on the last category as it provides a reliable model for SR and still knows advances in the field.

3.2.2.3 Simultaneous Bayesian approaches

Some frameworks allow to model the previous steps in a global optimization process. Notably, Bayesian frameworks have been proposed to consider the optimization of all the variables of the problem *i.e.* the registration parameters and the final high-resolution image. Examples of Bayesian approaches can be found in [HBA97, TB06, PCRZ06, PCRZ09, LS11, LS14]. Different frameworks have been progressively introduced, addressing more and more complex motion and interfering parameters. They can be described with the following derivation of the Bayes rule:

$$p\left(y_k \mid x, \Theta_k\right) = \frac{p\left(x\right) \times p\left(x \mid y_k, \Theta_k\right)}{p\left(y_k\right)} \tag{3.2}$$

where y_k is the k^{th} observation in a set of K_l LR images, x is the HR image to be recovered and Θ_k is a set of parameters that contains the SR observation model described in subsection 2.2.3, for each LR observation. A basic approach is to minimise the a posteriori log-likelihood, with known parameters Θ_k. In this case, we end up with a

classical linear problem with an unknown vector x, and a prior term $p(x)$. However, more recent approaches include the estimation of those parameters in the optimisation problem, which is the basis for a *blind* set-up that allows to recover not only the SR image but also the observation models parameters. While relatively simple subpixel motion is addressed in [PCRZ06, PRZ06, PCRZ09] with only translational motion and small rotation [TB06], [LS11, LS14] include an optical flow estimator in their framework, able to address more unconstrained motion. When adding those supplementary parameters, different methods can be chosen to solve the optimisation problem. One solution is to alternate between the parameters, by solving one while fixing the others. By marginalizing over some of the random variables, individual estimates can be obtained. [TB06] marginalize over the SR image to get the registration parameters. [PCRZ06] rather marginalize over the registration parameters, to use non-gaussian SR image priors such as the Huber prior. More accurate results can be obtained by taking care of outlier pixel and learning the priors parameters. As an example, [PRZ06] use a cross-validation mechanism over pixels in multiple frames to refine Huber prior parameters.

3.3 Single image SR

Considering the previous model of MISR, we focus on the case where $K_l = 1$; *i.e.* we only have one LR observation. In this case, only a fraction of the high-resolution information is captured. The inverse problem is therefore ill-posed and poorly conditioned, and a stable solution can only be achieved via coupled constraints on the low-resolution observation and external knowledge. This external knowledge can relax the problem, and analogue to the MISR case, supply priors that will condition the final SR image. We organise the review via four important areas in term of past works: Edge-Directed Interpolation, gradient profile prior, Bayesian approaches, example-based approaches. The latter will be more detailed as our work focuses on such methods.

3.3.1 Edge-directed interpolation

Although interpolation produces over-smoothed images, with noticeable artefacts (see subsection 2.2.2), it is still efficient for stationary or flat regions. An approach proposed in several works, referred as Edge-Directed Interpolation (EDI), aims to interpolate an image in a non-blind manner, *i.e.* taking into account the underlying shapes being interpolated. A meaningful option is to interpolate differently depending on the orientation of the gradient maps – as illustrated in Figure 3.2.

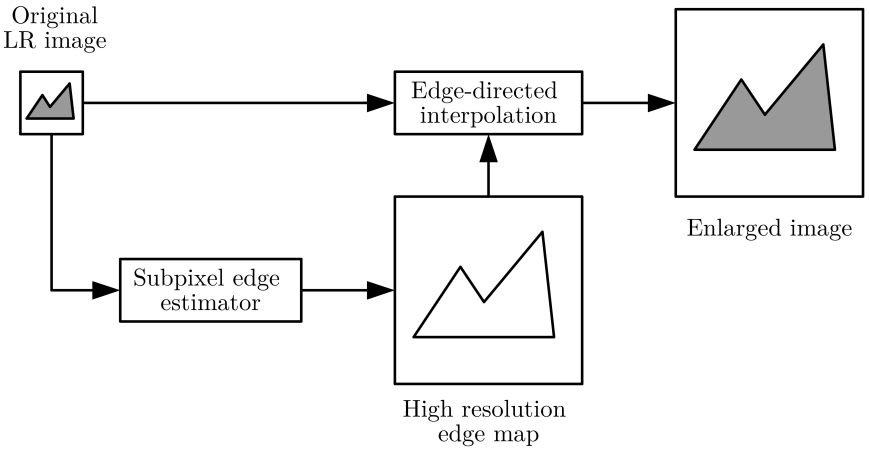


FIGURE 3.2: Illustration of EDI methods, figure reproduced from [AW96].

In [LP93], the authors propose an adaptive interpolation scheme based on zero-order (nearest neighbour) interpolation and b-spline interpolation. The algorithm is guided by the presence of edge, computed at each pixel as absolute differences with the neighbouring pixels, and utilised to classify each pixel into three groups of three edge patterns each. Depending on this classification, the value of the neighbouring pixels is not attributed following the edge direction. One of the first works using this approach was reported in [AW96]. They propose to generate high-resolution edge maps that allow to adapt the interpolation scheme depending on the presence of edges. The linear interpolation is therefore preferably performed perpendicularly to the gradient direction. Then, they use a constrained iterative procedure that alternates between high-resolution image generation using the described procedure, and low-resolution image refinement from this generated image. In [LO01], authors propose a similar approach, but model the interpolation scheme at edges using a relationship between the LR and the HR covariance.

3.3.2 Gradient profiles and natural priors

With the same intuition as EDI, several works propose to focus on edge regions in order to refine them, rather than adapting the interpolation process. They modify the gradient profile to produce sharper images as smooth edge is one of the fundamental artefact that appear when interpolating LR images. The gradient profile is defined as the section of the 3D gradient image in the direction of the gradient (or perpendicularly to the edge). The strength of those methods is their ability to address multiple scale using the same techniques.

In [Fat07], authors model the 1-D gradient profile as a continuity measure $\mathscr{C}$ (that can be seen as the intensity variation) defined via a conditioned Gaussian distribution. At

each pixel, a conditioning feature vector characterising its nearest edge and its context is computed. It is used to condition the distribution of $\mathscr{C}$ that will be used to construct the SR image with adapted gradients. Put together, this forms a Gauss-Markov Random Field from which the whole SR image can be sampled. They add a constraint (similar to the one described in [IP91]), that ensures that the HR image is coherent with the input LR image.

Another approach is proposed in [Sun08]. The authors model the gradient profile as a 1-D Generalized Gaussian Distribution. They define the more likely distribution's shape parameter λ using a set of natural images. This approach has the advantage to have less parameters than [Fat07]. They generate a high-resolution gradient profile map $\nabla\widetilde{x}^T$ by transforming the low-resolution one via this natural statistic. Then, the SR image is reconstructed via a gradient descent over the energy $E\left(\widetilde{x}|y, \nabla\widetilde{x}^T\right)$ that includes constraints on LR image reconstruction from the SR result and on the gradient map, defined as:

$$E\left(\widetilde{x}|y, \nabla\widetilde{x}^T\right) = \|y - DH\widetilde{x}\|_2^2 + \|\nabla\widetilde{x} - \nabla\widetilde{x}^T\|_2^2 \tag{3.3}$$

The authors further experiment the approach on blurry images in [SXS11], and perform sharpness transfer between images. They also report that the method is about five times faster.

A similar approach is taken in [TLBL10], and extended to better approximate high-resolution texture synthesis by the use of an "exemplar" input. This exemplar's gradient field is scaled down and up to form pairs of low and high resolution gradient fields. They are then included in the energy to be minimised (similar to equation 3.3) to generate adaptive textures and edges.

3.3.3 Example-based

Example-based methods make use of systems that have a learning and/or storage capability, and can perform a process given a set of training data. This global term can incorporate various type of approaches, from distance-based approaches to more complex ones, that model the underlying generative process, producing SR images conditioned to a given LR image. Since the 2000's, these methods have known growing interest in Super-Resolution. The reference papers [FPC00, FJP02] had a great impact on the exploration of such methods for SR. In this paper, the authors propose a nearest neighbour approach that consist in looking for the K_{nn} nearest neighbours of a given observed LR patch in a database of training samples, and select the best corresponding HR candidate via a markov network. They also propose an equivalent straightforward approach that directly selects the nearest neighbour via a robust patch representation.

In the following, we present the main research work that have been conducted for example-based SISR, divided into 4 categories: neighbour embedding and manifold learning (paragraph 3.3.3.1, sparse dictionary learning (paragraph 3.3.3.2, internal learning (paragraph 3.3.3.3 and neural networks (paragraph 3.3.3.4). Note that some methods may land in several categories (*e.g.* sparse code prediction using neural networks).

3.3.3.1 Manifold Learning and Neighbour embedding

In example-based SR, we can represent two data spaces which are a LR space and a HR space. Given a set of LR and HR samples, LR and HR manifolds can be inferred. However, the two manifolds do not necessarily exhibit the same shape, and two close samples in the LR space may result in distant samples in the HR space, due to the injective nature of the LR-HR relation.

The first method using such approach can be found in [CYX04]. Instead of selecting a single candidate for the HR patch, the authors make the assumption that LR patches can be represented as a linear combination of their K_{nn} nearest neighbours. Furthermore, they consider that the LR and HR manifolds are locally similar, and that the corresponding HR patch can be found using the same linear combination of the HR versions of the found LR nearest neighbours. They propose to use extract features using first and second order derivatives, and set $K_{nn} = 5$. The nearest neighbours are found using euclidean distance, and the weights of the linear combination of the K_{nn} LR patches are found using matrix inversion. The results are less noisy than [FJP02], but also smoother due to the averaging effect of the linear combination. This makes the images look less realistic in some regions.

To overcome those limitations and reduce the computational cost, other approaches have been proposed. For instance, in [BRGM12] the authors propose to work with different patch features (normalised luminance and derivatives) and to use a non-negative constraint in the LR neighbourhood computation. They show that their approach allows to work with more coherent manifolds, as the non-negative weights computed from the LR embeddings allow to reconstruct the HR patches better. In [GZTL12], the authors propose to extract LR features using Histogram of Gradients and use K-means to create clusters of similar histograms. More recently, two State-of-the-art approaches – namely Anchored Neighbour Regression (ANR) and Adjust ANR (A+) – have been proposed, based on the dictionary construction from [ZEP10] but using different approaches for sparse search and regression through the dictionary atoms. The ANR approach [TDG13] is based upon the following principle: a patch is likely to be a linear combination of atoms that are close from each other on the hypersphere they lie on. Therefore, the search for a

sparse vector α in the dictionary is replaced with i) the search of the nearest atom from the processed patch and ii) an offline-calculated projection matrix allowing to produce the corresponding HR patch. The search is a simple correlation between the features extracted from the processed patch. The projection matrix is calculated from the K_{na} nearest atoms of each atom. The method can be summarised as follows:

1. Build a dictionary using K-SVD and OMP [AEB06, RZE08], similarly to [ZEP10]

2. For each atom d_j of the dictionary, select the K_{na} nearest atoms, and compute the projection matrix P_j:
$$P_j = D_h \left(D_l^T D_l + \lambda I\right)^{-1} D_l^T \tag{3.4}$$
where D_h is the HR dictionary, D_l the LR one and λ is the regularisation parameter

3. at test time, compute a high resolution patch x_n from a low resolution one y_n using:
$$x_n = P_j y_n \tag{3.5}$$
P_j being the projection matrix of atom d_j so that:
$$d_j = \min_{d_i \in D_l} \langle y_n, d_i \rangle \tag{3.6}$$

In [TDSVG14], the authors use the same principle of neighbourhood in the atom space, but this time compute for each atom a linear projection computed on a pool of K_{ns} training samples that lie close to a given atom, instead of the previous projection matrix. This change allows to capture the trends in the neighbourhood directly in the data but with a fine selection thanks to the atoms. The projection matrix being calculated offline, there is no time consumed with this operations.

1. Build a dictionary using K-SVD and OMP [AEB06, RZE08], similarly to [ZEP10]

2. For each atom d_j of the dictionary, select the K_{ns} nearest pairs of examples from the training set (S_l, S_h), and compute the projection matrix P_j:
$$P_j = S_h \left(S_l^T S_l + \lambda I\right)^{-1} S_l^T \tag{3.7}$$
with S_l being $K_{ns} \times D$

3. At test time, compute a high resolution patch x_n from a low resolution one y_n using:

$$x_n = R_j y_n \tag{3.8}$$

with

$$d_j = \min_{d_i \in D_l} \langle y_n, d_i \rangle \tag{3.9}$$

Semantic context matching Several methods use patch clustering, which makes sense as LR basic structure are still likely to have similar shape in the HR domain. However, at a higher semantic level, the similarities can be helpful as well to reduce the complexity of the manifolds. Intuitively, if we can preselect images with the same content as our LR observation (*e.g.* forest, buildings, animals), we are more likely to obtain coherent reconstructed SR images. In [SZT10], the proposed approach makes use of the Berkley Segmentation Engine (BSE) to first produce images segment that have a uniform texture. Similar candidate patches are then extracted from those segments depending on the similarity of the distribution of texture features, using an approximation of the Kullback-Leibler (KL) divergence distance. The result of this search is illustrated on Figure 3.3. The fusion of the candidates is obtained *via* a three-termed optimiza-

FIGURE 3.3: Result of similar segments using the BSE segmentation of natural images and texture similarity search with the KL divergence, illustration from [SZT10]. The left column represent input images and the two other columns the more likely candidates found by the search.

tion problem including a reconstruction term (coherence with the LR observation), a Hallucination term (the correspondence of the SR image with the candidates) and an Edge Smoothness term. This idea is extended in [SH12] where the search is done via a large-scale search over the Internet, to find images similar to the LR observation.

In the following, we describe Sparse dictionary learning, where dictionaries of low and high resolution samples are constructed to perform SR. Note that some approaches are hard to classify between Neighbour Embedding and Sparse Dictionary learning, as the notion of "neighbourhood" can be common in both.

3.3.3.2 Sparse Dictionary learning

The main idea behind dictionary learning for SR is to create a set of so-called atoms for which we possess a LR patch and its HR counterpart. The most simple and intuitive way to create an atom (or dictionary element/class) is to take a – possibly normalised – sample pair. When given a LR image, one can simply split it into LR patches, find a representation using the LR dictionary atoms and use the corresponding HR atoms to reconstruct the HR image. Through the years, efficient ways of creating dictionaries have been proposed.

The first main contribution to the use of dictionary was made in [YWMH08]. The authors propose to use an overcomplete sparse dictionary of low and high resolution image patches (about $100,000$) to perform SR. A given LR patch y_i is represented as a sparse linear combination of LR dictionary elements αD_l, and a HR patch x_i is reconstructed using the corresponding HR elements $D_h\alpha$.

$$y_i = D_l\alpha_i \tag{3.10}$$

$$x_i = D_h\alpha_i \tag{3.11}$$

For each patch y_i of an image, an optimal α_i can be found to minimise the following equation:

$$\underset{\alpha_i}{\operatorname{argmin}}\|y_i - D_l\alpha_i\|_2^2 + \lambda\|\alpha_i\|_1 \tag{3.12}$$

where λ balances the role of the sparsity of α_i. The authors propose to add a constraint on the top and left border region of the reconstructed HR patch $PD_h\alpha_i$ to ensure its compatibility with the previously reconstructed image overlapping the region, w:

$$\|w - D_h\alpha_i\|_2^2 \tag{3.13}$$

Put together with 3.10, we have:

$$\underset{\alpha_i}{\operatorname{argmin}}\|\widetilde{y_i} - \widetilde{D}\alpha_i\|_2^2 + \lambda\|\alpha_i\|_1 \tag{3.14}$$

with $\widetilde{y_i}$ and $\widetilde{D}$ including both constraints. The sparsity is ensured via the use of a a L_1 norm constraint on the sparse code α during the optimization process. They also use a global reconstruction constraint using the backprojection method, where the consistency of the high-resolution image with respect to the model and the low-resolution observation is ensured. This approach is further improved in [YWHM10] where more compact dictionaries are constructed from the extracted patch pairs, using an iterative sparse coding optimisation algorithm [LBRN06].

The authors propose to jointly learn two dictionaries to perform SR: one in the LR space and one in the HR one. To learn such dictionaries, they extract corresponding patches from low-resolution (bicubic) and high-resolution images, and add the constraint that the LR and the HR patch can be linearly reconstructed from their respective dictionaries using the same sparse vector α. During optimization, the algorithm alternates between the construction of the dictionary (with a fixed sparse code) and the search for a sparse code (with a fixed joint dictionary). This gives the following minimisation objective during the construction:

$$\underset{\alpha_i, D_l, D_h}{\operatorname{argmin}} \|y_i - \widetilde{D}\alpha_i\|_2^2 + \lambda\|\alpha_i\|_1 \tag{3.15}$$

For the reconstruction, D_l and D_h have been learned and are therefore fixed, and only the Z estimation step is required.

Extensions and improvements Several works have proposed improvements from this sparse formulation. In [DZSW11], the authors explore subdictionary learning for both deblurring and SR. Instead of learning a single dictionary, they first partition the high-pass filtered image space using K-means. Then, for each cluster, a dictionary is learned using a L_1 constraint on the sparse code, resolved with an iterative shrinkage algorithm [DDDM04].

The approach has also been improved in [ZEP10], with the ability to fine-tune the dictionary with respect to the current image, taking it as a HR image. However, they use K-SVD to encode the LR sparse representation and use a pseudo inverse to recover the high-resolution representation. The sparse coding is realized using Orthogonal matching pursuit (OMP), rather than solving the LASSO optimisation problem with constraints on the sparse code. They use 1000 atoms as dictionary size for the LR representation. This approach is sped up in [YWL+12] using a neural network for the sparse vector inference, which becomes a forward pass instead of a search optimization algorithm.

In [WZLP12], authors propose a different coupling between the sparse codes, which is a simple linear relation between the LR and HR sparse representation. It can be seen as

a coupled dictionary of LR patch and HR sparse codes instead of using the same sparse representation.

3.3.3.3 Internal Learning

Several approaches make use of internal patch recurrence across scales and spatial dimensions to avoid external database dependency. This means a pool of low and high resolution images is extracted directly from the observed image. This supposes the observed images are sufficiently large to contain interesting and various content that can be observed at different scales. Figure 3.4 summarises this principle.

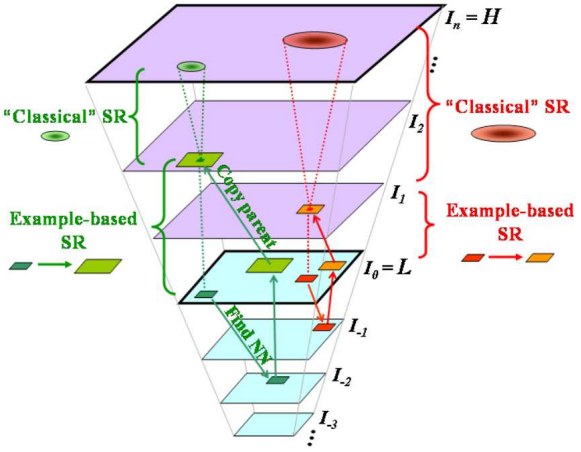


FIGURE 3.4: Internal learning approaches capitalise from multiscale analysis of the input image to create in-place example pairs for learning. This illustration is extracted from [GBI09].

To perform this internal learning, the idea of adapting the Non-Local Means (NLM) algorithm [BCM05] have been explored in [EV07] and [PETM09]. The NLM algorithm is a non-local patch-based method that, for every position in the image, looks for similar content – in terms of L_2 norm – in a neighbourhood (i.e. non-locally), and perform a weighted sum depending on the similarity. In [EV07], the authors point out the missing link between the example-based, non-local approaches and the pyramidal or multi-resolution approaches such as fractal ones. They propose to use a generalisation of the NLM algorithm to perform super-resolution. The generalisation includes external example images for similarity search and across-scale search. A similar approach is presented in [PETM09], with complementary experiment and extensions. First, they fuse zero-padded version of the LR similar non-local patches. They propose to use a deblurring step, as the adapted NLM procedure mainly compensates for the noise and the downsampling process, but to a lesser extent for the blur in the LR image synthesis

model. Second, they extend the approach to multiple-image, which is slightly different from the external image proposed by [EV07] as they correspond to low-resolution observation. The NLM approach is equivalent to an implicit, fuzzy motion estimation.

Apart from NLM algorithm, other approaches have been proposed. In [GBI09], patch cross-scale similarities are exploited: at each position of the input LR image, a patch is extracted and its nearest neighbours are found in the lower scales. Then, their upper scales counterparts are used as HR examples for the current processed patch. They use small scale steps (1.25 or $2^{2/3}$) to adopt a coarse-to-fine iterative algorithm which is stable through the experiments. The consistency of the SR image at *each* scale is checked using backprojection towards the appropriate scale. In [YHY10], the authors use the same approach but construct a dictionary by clustering similar patches, and finally use it to perform SR. Contrary to the previous presented works, authors in [FF11] propose to use a local upscaling scheme. They use the shape redondancy between a $(M+1)\times(M+1)$ patch and its downscaled version of size $M \times M$, when the selected scale is low. They use $M+1:M$ scale steps. To perform those small scale steps, they use non-dyadic, biorthogonal filter banks. In [HS11], a Gaussian Process Regression is performed, with extracting training patches in the bicubic-interpolated image and the LR one. The framework allows to simultaneously deblur and produce sharp images at higher resolutions. Moreover, instead of looking for similar patches in the image domain, the algorithm proposed in [SA14] make this search in several sub-band images, obtained by oriented bandpass filters, which results in lower reconstruction errors. To account for the relative rareness of in place examples compared with large-scale external learning datasets, in [HSA15] the authors use the PatchMatch algorithm [BSFG09] to enrich the search of non-local patches in the image. This allows to search not only via 2D translation (horizontal and vertical dimensions in the image) but also via linear transformation (scale, rotation, perspective), yielding a 7-dimension search, comprising a plane index, where planes are estimated to account for the perspective effect that occur in natural images.

3.3.3.4 Neural Networks-based SR

Neural networks has known big advances with the recent trends in deep learning. Regression problems can be addressed via Neural Networks, for which an euclidean loss between the output of a neural network and the targeted signal may be used.

Primary/Previous work on Super-Resolution using Neural Networks In [AGK95], a radial basis neural network is used to perform interpolation. Each centre of the hidden layer is initialized with a training example and the output layer performs a linear

regression to predict the value of the central pixel. A variable variance parameter on the gaussian radial function allows to adjust the sharpness of the resulting image. The experiment is conducted on a single 20×20 image and does not provide objective evaluation as no reference image is used for the higher resolution.

In [Pla99], authors propose to predict the missing pixels for the line doubling and interpolation problems. They use a one-hidden layered MLP to predict the graylevel values. For the interpolation problem, they use 24 pixels as an input patch, 16 hidden neurons and 5 output neurons, as described in Figure 3.5 (from the original paper). They reported good results and a robustness against Gaussian and coding noises.

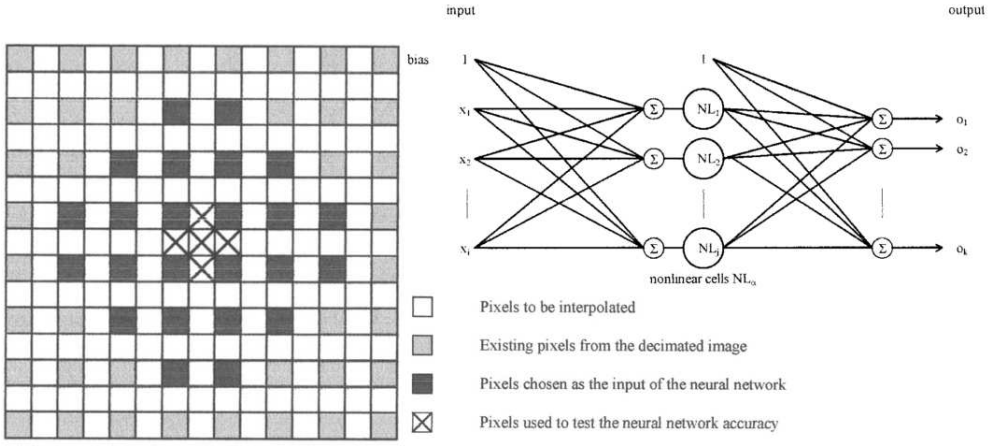


FIGURE 3.5: Methods proposed in [Pla99] for interpolation using a one-hidden-layer perceptron Neural Network. The figure is a compilation from those presented in their paper.

Similarly, in [DH00], the author propose to use a 1-hidden layered MLP to predict the values of the missing pixels in an HR grid for binary images. The setup is similar to [Pla99] but works on binary input images and give more insight about the limitations and the relation between the training data and the observed improvment in PSNR. A similar approach is taken in [GSL00], but adapted for Bayer color images. A MLP is learned for each colour channel with different selection for the 4×4 input patch, depending on the density of color sensor (2 times more for green than for blue and red).

Later in [ZP02], the authors propose to learn the optimal mapping between LR and HR residuals. The LR residuals are obtained as the difference between a downsampled interpolated LR image and the original LR image while the HR residuals are the difference between the HR and the LR image. They give detailed results on the Kodak dataset for $\times 2$ SR and compare with [Pla99] and [AGK95] approaches, reporting better results in terms or PSNR .

In [Kum03, LFP+07], fuzzy approaches are proposed to account for the context of the processed patch. On smooth regions, a fixed linear interpolation is performed while an adaptive neural network predicts the weights of a variable linear interpolation kernel on non-smooth regions. In this paper, the authors capitalize on Kondo's method which clusters data and uses a linear filter for each cluster. They propose to learn a non-linear filter using MLP instead of a linear one. The input 3×3 image patches are classified according to Adaptive Dynamic Range Coding (ADRC) scheme, which is similar to a local binary pattern (LBP) one but with an inequality threshold defined by the mean of the 9 values:

$$i_{ADRC} = \begin{cases} 0 \,\text{if}\, i < i_{average} \\ 1 \,\text{if}\, i \geqslant i_{average} \end{cases} \tag{3.16}$$

instead of the central value for LBP:

$$i_{LBP} = \begin{cases} 0 \,\text{if}\, i < i_{central} \\ 1 \,\text{if}\, i \geqslant i_{central} \end{cases} \tag{3.17}$$

They use a larger training database composed of $200,000$ patches in each class, constituting one of the first large dataset for training a neural network for super-resolution.

Autoencoders With the recent work on autoencoding architectures, different approaches were proposed to take advantage of autoencoders for SR.

For instance, Restricted Boltzmann Machines (RBM) are used in [GGY13]. Their approach is strongly related to sparse coding methods. It encodes a dictionary of LR/HR patch pairs in a RBM and take advantage of the framework to iteratively reconstruct an HR image as a sparse mixture of the embedded patches, via a sparse activation of hidden neurons. They obtain similar results to neighbour embedding [CYX04] and sparse coding [YWHM10] methods but with increased speed and an elegant framework. RBM are also used in [PE14] to encode a relationship between sparse representations in overcomplete dictionaries.

[NTA13] makes use of a Deep Belief Network (DBN) to learn the autoencoding of the DCT coefficients of HR images. Then, from the low-frequency coefficients of a scaled-up LR image, the network iteratively recovers high-frequency as it is the only kind of image it has learned to produce/generate. In the image domain, [CCS+14] propose a Deep Network Cascade (DNC) that gradually upscales an input image, until the desired resolution has been reached. They employ a non-local self-similarity search to find cross-scale examples that match with the processed patch, with a back-projection constraint.

In a similar fashion as [DLHT14], authors from [WYW+15] use an auto-encoding convolutional structure. The encoder is composed of several convolutions and the decoder is a set of "deconvolution" layers, that consist of zero-padded convolutions. They show that this approach allows to learn more localised filters. This is due to the fact that borders have to be reconstructed using the available spatial data, and not only the centre of the processed region. They also perform online adaptation of their model using data augmentation and fine tuning. Given a single image and a pre-trained model, they create pairs of low and high resolution patches similarly to the internal learning approaches. The data is augmented by using slightly higher and lower scaling factors (1.2, 0.8). They demonstrate that content adaptation can further increase the effectiveness of learning-based approaches.

Neural Network-based sparse coding In [OSS14], the authors propose to use a CNN to approximate the sparse coding methods. Inspired from the work presented in [KSB+10], they account for the necessary resampling for resolution change. Instead of upsampling the original LR image, they upsample the sparse feature maps by adding a linear upsampling matrix in the formulation of [KSB+10] in the decoder part. The learning process is done via alternating between Fast Iterative Shrinkage-Thresholding Algorithm (FISTA [BT09]) to approximate the sparse code and one step of stochastic gradient descent to update the convolution filters. They show results that improve over conventional interpolation technique while staying competitively fast, as they only use 8 maps.

In [WLY+15], a sparse coding network is also proposed but taking the bicubic image as an input. Their network is mimicking the sparse coding approach while using Learned Iterative Shrinkage-Thresholding Algorithm (LISTA [GL10]) and SGD: a LR patch is extracted with fine-tuned Haar-like extractors and fed into recurrent LISTA stages that iteratively approximate a sparse code. Finally, the sparse code is used to recover HR patches that are reorganised with a last convolutional layer.

Convolutional Neural Networks Convolutional Neural Networks have drown a lot of attention in the last decade. While several works reported competitive results since the 90's [LBBH98], new paradigms in computation, data handling and optimization algorithms have allowed breakthrough in many signal processing applications, particularly in image processing and computer vision area. These architecture play a major role in Deep Learning as they allow to extract multi-scale hierarchical features from the processed signals. While previous work include the use of MLP, we now present recent advances in SR that take advantage of the modern ANN design.

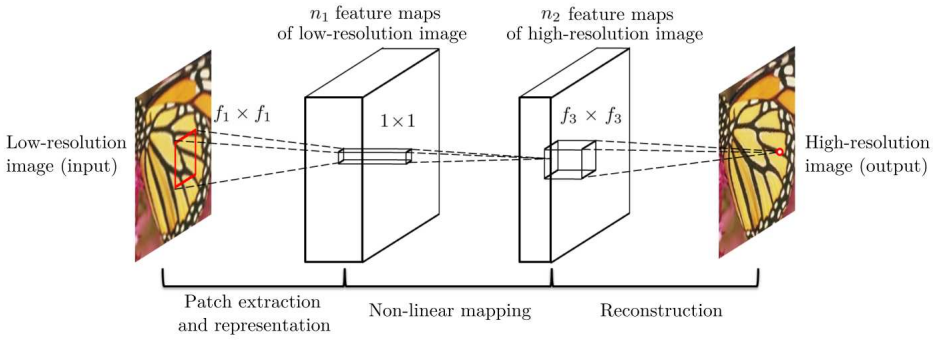


FIGURE 3.6: Standard CNN architecture for SR: the upsampled image (using bicubic interpolation for instance) is convolved by learned non-linear filters in different layers, merged to form the final High-resolution image. The illustration is taken from [DLHT14].

In [JMR+07], authors were the first up to our knowledge to use a CNN approach to perform super-resolution, called *super-sampling restoration* in the paper. Their goal is to use larger images to obtain better segmentation and binarization performance. They use several filters that are reorganized spatially, and that can be used in the 3D case (2 spatial dimensions and one temporal).

Later in [DLHT14], the authors proposed to use a three-layered convolutional neural network to perform super-resolution. Starting from the bicubic-interpolated image, they train the network to predict the grayscale HR image. The network serves as a feature extractor in the first layer ($64 \times 9 \times 9$ convolution filters), and maps the features to a HR space in the second layer ($32 \times 1 \times 1$ convolution filters). The third output layer acts like a final mapping between the features and the SR image ($1 \times 5 \times 5$ convolution filters). This elegant framework allows to avoid the choice of features as it learns its own. It can be further argued that the learned features are dedicated to the specific task of SR. The reconstruction of the grayscale image also gives the approach a generative flavour as the whole signal is reconstructed. This aspect is however limited as the input signal is rather filtered than encoded in the network. The model and the choice of convolution sizes also compare their approach to a dictionary one with feature extraction, LR-HR mapping and final averaging and spatial coherency. In [DLHT16] the same authors explore further configurations where this analogy with dictionary learning is less explicit. They reach slightly higher scores. Their experiments show that using four layers does not increase the performances of the network. They also use two different datasets for training their network. The first one is the same as [YWHM10] (91 images, yielding $24,800$ subimages of size 33×33) ; the second one is a subset of Imagenet [RDS+15] containing $395,909$ images, from which they extract 5 million 33×33 subimages. They show that using a large scale dataset allows to gain over smaller dataset, even though the model does not hold many parameters.

In [KLL15a] the authors propose several improvements over SRCNN [DLHT14]. First, like in [ZP02] they use high-frequency targets and not the whole signal. They also saturate the back-propagated gradient using gradient clipping to benefit from higher learning rates while avoiding exploding gradients. They show that better results can be obtained. They also augment the data to take several scales into account. The results indicate that a single model can handle several levels of bicubic decimation which is a form of blind set-up for Super-Resolution.

In parallel, they also proposed a second architecture [KLL15b] in which the multiple layers of the deep network are replaced with a recursive convolutional layer called *inference network* that iteratively refines the predicted maps and gathers more spatial context by its convolutional nature. As shown in Figure 3.7, the architecture is composed of an embedding network that projects the input image in a suitable space, followed by the described *inference network*. At the end, a *Reconstruction network* maps the obtained deep features to the image space. They experiment different strategies to overcome training difficulty that arise due to the depth and the recursive nature of the network. They find that a combination of *skip-connection* (*i.e.* using intermediate output of the recurrent layer) allows a better training and good performance.

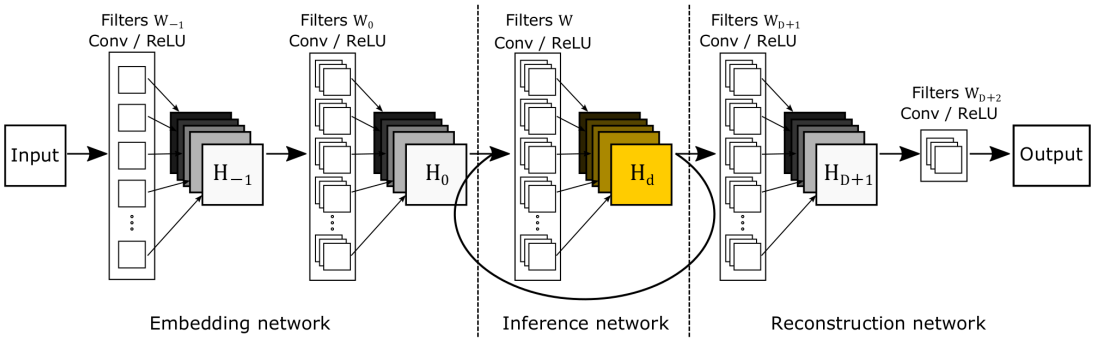


FIGURE 3.7: Using a "recursive" layer and "skip-connections" tricks allows to gather more spatial context with the same number of parameters while using intermediate representations to predict the final SR image. The illustration is taken from [KLL15b].

In [YFY+16], the authors also propose a deep convolutional architecture that aims at reconstructing images and edges. The input data is composed of the interpolated low-resolution image stacked with edge maps, extracted with simple hand-crafted edge detectors. To account for the fact that desired details to be recovered belong to different frequency band, they propose to iteratively estimate the higher bands from the lower ones. This iterative nature is embedded in the network using recurrent layers that predict residuals of the higher bands.

All the neural based approaches above are trained to minimise a pixel-wise error between the SR image and the original HR one. This makes the prediction difficult for areas with

complex spatial statistics. In the following are reported several works seeking to produce credible high-resolution images by incorporating long-range consistency and high-level knowledge into the training process.

Perception-based SR Traditional learning-based algorithms being trained to minimise the difference with the HR training image are good on structural, highly erroneous regions such as edges. However, they tend to produce oversmoothed results on regions that would contain fine-grain details and textures on realistic images. These textures are too diverse and complex to be learned from pixel-wise error such as MSE with a high-resolution image. Perception-based SR approaches are a very recent set of methods that aim to take the visual aspect into account, which is a reasonable angle as human perception of HR image does not depend on a pixel-wise distance but rather on a large spatial scale coherency. To expand the example of fine-grain textures, those approaches would sample a coherent, high-resolution texture (*e.g.* grass, bricks, skin...) rather than trying to recover the original texture.

In [BSL16], the authors propose to generate SR image using deep models for feature extraction, allowing a stronger abstraction towards the data. Two deep structures $\Phi(x)$ and $\Psi(y)$ working on the LR image x and the HR one y are coupled via a L_2 norm on their feature vectors. The first extract a vector of features from the LR space. The second is trained to extract features from the HR image and exhibits a revertible nature, allowing to update the input data given a desired output feature vector. This work can be seen as an interesting alternative to the pixel-wise losses used in [DLHT14], as it brings general purposes features in the SR game. More precisely, the Ψ network is either a pre-trained CNN based on the VGG-19 architecture [SZ14] or a wavelet-based scattering network based on [BM13]. The second network Φ is a 5-layer CNN that is trained with pairs of LR images x_i and corresponding HR features $\Psi(y_i)$.

In the same spirit of bringing perceptual knowledge to SR, the authors in [JAFF16] propose a so-called transform network, and use it for style transfer and Super-Resolution. The concept is to train a network to reconstruct an image that holds not only good low-level (pixel-level) aspect but also desirable large scale semantic content. To achieve this, they use several low and high-level losses functions as depicted in Figure 3.8. The low-level loss function denoted ℓ_{pixel} are classical Euclidean and Total Variation (TV) ones, which contribute to render images with natural properties and similar to the HR image (in the case of SR). The high-level loss functions are euclidean-based distances between higher-level neural features ℓ_{feat}, obtained through deep convolutional neural networks (VGG16 [SZ14]) trained on large scale recognition datasets. A normalised L_2 norm can be computed to translate the difference between two convolution maps, one for

the training image and on for the reference image. For Super-Resolution, they separate the use of pixel-wise losses (ℓ_{pixel}) and semantic losses (ℓ_{feat}). They also use histogram matching between the generated SR image and the input LR image as a post-processing step. The reported results show that the ℓ_{feat} loss yields images with pleasing visual content: long-term coherency between the texture and objects, sharp and continuous edges. As expected, it performs less good than pixel-wise trained SRCNN architecture in term of PSNR and SSIM, which are low-level measures. The ℓ_{pixel} loss does not give competitive performance compared with SRCNN, but is still better than bicubic upscaling.

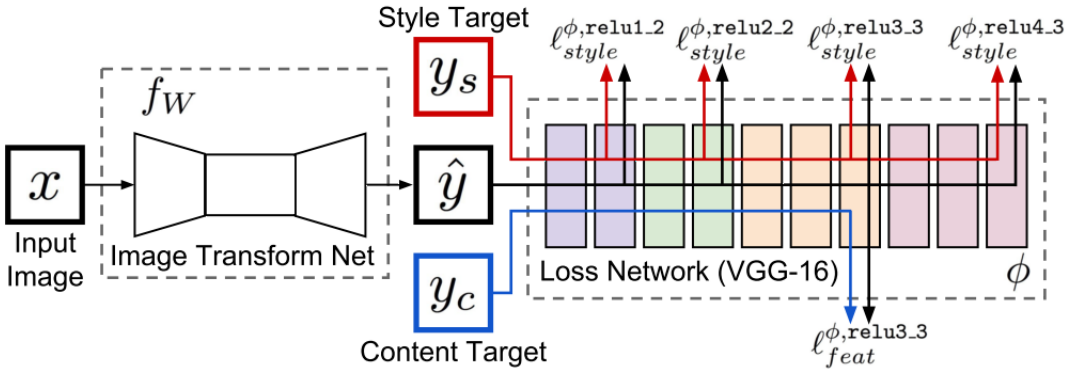


FIGURE 3.8: Perceptual Losses allow to bring cost functions that depend on high-level representation of the output image in the learning process, instead of the usual pixel-wise squared error with the high-resolution image. The figure is taken from [JAFF16].

More recently, inspired by recent work on Generative Adversarial Neural Networks (GANs) [GPAM+14], the authors of [LTH+16] proposed a SR scheme with realistic rendering. GANs make use of two experts that compete with each other (see Figure 3.9). For SR, one expert generates a SR image (conditioned to a given LR image), and another expert classifies it into either a HR natural image or a SR image. The generative model is a CNN composed of a non-linear convolutional layer, followed by 15 identical residual blocks that advantageously gather more context using a single set of parameters. Finally, two "deconvolution" layers (*i.e.* convolution with zero-padding) allow to reconstruct the SR image. The discriminator is a VGG-like network, that uses both convolutions and strided convolutions (applied every other position) to reduce the spatial dimension and outputs a single sigmoidal output that classifies the input image into a real or a SR image. They propose several losses to address the SR task. As in [JAFF16], and using the residual architecture from [KLL15b] they first experiment with low-level loss: the classical MSE is used to train a so-called SRResNet. Second, they propose a perceptual loss composed of three terms: a *VGG loss* similar to [JAFF16], an adversarial loss guided by the output probabilities of the classifier, and a *regularization loss* based on total variation of the output SR image that favours smoother solutions by limiting the spatial intensity changes. The results of the SRResNet have high PSNR

and SSIM scores but yield oversmoothed images, while the images produced by the adversarial architecture have highly likely and realistic visual shapes and textures.

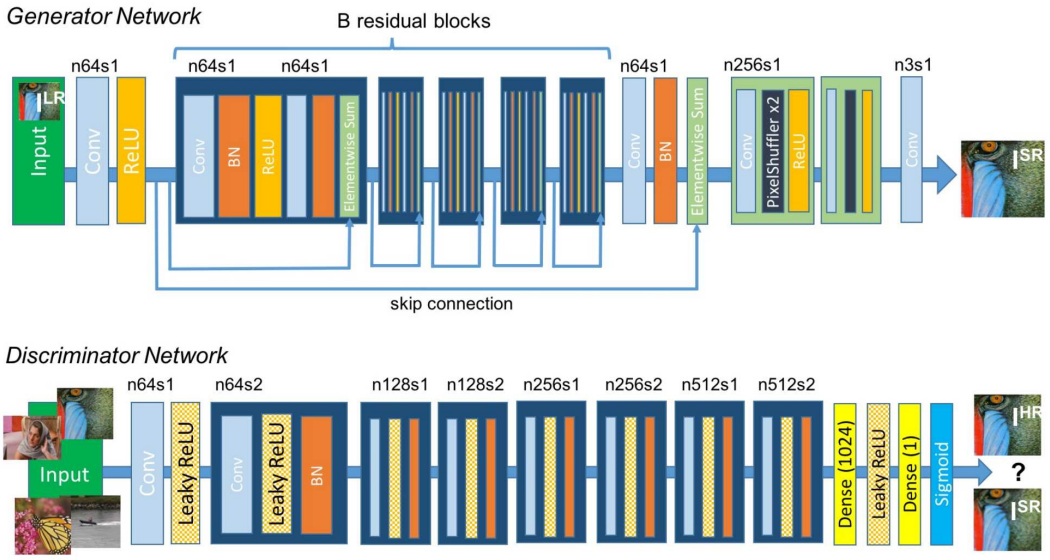


FIGURE 3.9: In the GAN approach proposed in [LTH+16], a generator and a discriminator are competing to respectively produce more and more realistic SR images and distinguish more and more accurately between SR images and HR images. The more the discriminator network can tell the difference between the two, the more the generator network is challenged to produce more confusing (thus realistic) images.

3.4 Domain-specific Super-Resolution

As mentioned in subsection 2.3, SR can broadly be separated into two groups of applications: general purpose applications for which the goal is to perform well on generic images, mainly for visual enhancement (see paragraphs 2.3.1 and 2.3.2); and domain-specific applications (see paragraph 2.3.3), where a strong a priori is assumed on the content of the image and on the goal is rather to allow a better interpretation – either by a human or a machine. In this section, the literature of text images SR and facial SR is reviewed, as some contributions of this thesis address these two domains (see chapters 4 and 5). Low-resolution texts are likely to appear in many situations (distance to the camera, tiny fonts in a scanned documents, etc.), and improving their resolution can help human readers or OCR systems that require high resolution input images. Dedicated SR methods can use priors that are specific to text images (*e.g.* foreground/background relationship), or use non-local assumptions such as the presence of repetitive patterns (characters, words). They are reviewed in the next subsection 3.4.1. Facial (or face) image SR is also referred to as face hallucination. Indeed, those SR algorithms have a strong *a priori* on the presence of a face and can "hallucinate" high-resolution faces

features (eyes, nose, mouth) inside LR images. Different ways of incorporating these aspects are reviewed in subsection 3.4.2.

3.4.1 SR of Textual Images

Several works focus on textual images. Different assumption can lead the choice of a method. Particularly, some methods are developed for document image, *i.e.* images obtained with the digitalization of a document (most of the time, black text over a white background), using different devices (scanner, hand-hold device). Other methods focus on license plates or other sources of text image such as TV or text "in the wild". The interest is not only to produce sharper image but also to increase the readability of the text. A richer evaluation is therefore possible, associating a pixel-wise measure with an objective measure – relatively to a chosen OCR system.

With this perspective, many works have been conducted, with an interesting diversity of approaches that we illustrate in the following. A recent review on text image SR may also be found in [WDL+16]. In [LKD99], authors propose text image enhancement techniques with a focus on registration and driven by text-specific problematics (horizontal text, background complexity, resolution). Their approach produces $\times 2$ bigger images as they proceed with subpixel registration (using interpolated images). A more SR-driven method was proposed in [CZ00], authors review the IBP algorithm and propose two estimators for text image sequences: a Bayesian MAP with a Huber prior, and a TV-regularized one. They first show that a simple ML estimator has low robustness when solved by optimization, while IBP performs better thanks to the choice of back-propagation function. Then, the two proposed regularized approaches are described and evaluated. They demonstrate higher robustness in producing piecewise images (desirable for text image: background and foreground) and finer shaped letters. Implementation of gradient descent optimization for SR is also well detailed. A non-linear optimization framework is proposed in [TC00], to address low-resolution document image SR. Three "scoring functions" are defined ; namely Bimodal, Smoothness and Average (BSA) scores. The bimodal score ensures that the pixel intensities exhibit a foreground (black) / background (white) statistical behaviour, corresponding to text and background. The smoothness score helps regularizing the spatial coherency and the average constraint score ensures that the obtained image is compliant with the original LR image. The derivatives are analytically obtained for each score, which allows a tractable iterative resolution. The methods yields better OCR accuracy over a large set of documents, compared with the spline interpolation. Another approach is presented in [DM05] where the authors explicitly use a multi-frame Bayesian framework similar

to [CZ00] but add a text-specific bimodal prior. This prior add a supplementary constraint on the image that accounts for the bimodal (foreground/background) nature of document text images. It is modelled as a bimodal Gaussian distribution, where the two modes are calculated from the SR estimate using an Expectation-Maximization (EM) algorithm continuously with the optimization algorithm (gradient descent). They test several combinations with a Huber smoothness prior on the gradient of the estimated HR image. The best parameters are manually selected to yield the best SR text image.

Another MISR method proposed in [MTM05] uses the Teager filter [MS01] instead of a specific prior. This filter allows to extract the high-frequency from the LR images while being robust to noise. Using image registration based on first order Taylor serie expansion, they fuse the original LR frames and the HR frames separately, using outlier frame rejection. The two obtained images are interpolated using linear interpolation and summed up to for the SR image, which is denoised using a 3×3 spatial median filter.

Taking another perspective, the authors in [LP07] perform a non-local search to take profit of the repetitive nature of single text images. As letters are likely to appear several times under slightly different conditions in LR document images, the non-local search allows to gather the similar shapes and fuse their small variations to produce well-shaped SR image. The fusion is performed using a median approach, less sensitive to outliers, and a different scheme is proposed at pixel level on the HR grid, depending on the presence of original, fused or unknown pixel values. The approach is followed by a denoising and deblurring step using total variation. The philosophy is therefore close to MISR (see section 3.2) and internal learning, non-local approaches (see paragraph 3.3.3.3), but is especially relevant for document images.

Sparse coding has also been applied to single text image SR. In [WDL+13, WDA+14], the authors use multiple dictionaries to perform SR on document and handwritten text images. The training patches are first partitioned into C clusters in an unsupervised fashion using the intelligent K-means algorithm proposed in [Mir05]. Then, a joint dictionary learning takes place in each cluster. At test time, a input patch is used to recover the sparse joint representation which is used to reconstruct the HR patch in each cluster. The cluster leading to the best LR reconstruction is chosen among all the C clusters [WDL+13] or just the K_{na} nearest ones [WDA+14] – determined by the euclidean distance between the input patch and the centroids – for the initial reconstruction. Then, the overlapping patches are averaged. To avoid artefacts, a IBP step is performed to ensure HR-LR compatibility, and bilateral filtering is used to remove ringing artefact while preserving edges. The obtained image are cleaner than other sparse coding methods and improve readability.

3.4.2 SR of Facial images

The community working on facial images have developed many SR methods, rather known as Face Hallucination methods. The term "hallucination" can be understood by the fact that those methods often have a strong a priori on the content of the image ("there is a face", "all faces are aligned", etc.). Thus, in most methods, the location of facial components is crucial and faces are often aligned. From there, some "global" methods consider the whole face while other "local" methods focus on fixed locations (*e.g.* on a regular grid) that will more likely contain the same content for each face. The first ones will capture the relationship between components but require an approach able to capture all the variations without loosing the particularity (*i.e.* for face recognition usage). The second ones have a more concise role but run parallel and independent SR "Experts" on each location and may miss coherency on the overall face (identity, luminance and contrasts). They are often regularized to yield a likely global face. Some other methods are more dynamic and make use of facial components detector to enhance each component independently. These approaches are more suitable for generic systems where the faces are potentially not aligned.

3.4.2.1 Global approaches

One of the leading works of the domain was conducted by [BK00], where authors report performance of several algorithms and propose a novel MAP approach with a prior based on gradient pyramids, specific to aligned faces. Solving the global MAP image by gradient descent, they demonstrate the ability of such algorithm to produce high-resolution images, even though artefacts are present in the final images. The algorithm is suitable for both single image and multiple-image SR. They also conduct extensive experiments and comparisons to study the impact of scale factor, noise, warping and occlusion.

Instead of images pyramids, other approaches use linear projections such as PCA to capture the global features of the face. A two-step approach is proposed in [LSZ01, LSF07]: the low-resolution facial images are first projected in a higher resolution space using a global parametric model, learned by PCA. Then, this first estimate is enhanced using a local, patch-based non parametric model based on a Markov network. The latter assumes that overlapping patches form a network that incorporates an internal potential function (the compatibility function with adjacent patches) and an external one (all the training patches at this position). Thus, it assumes a good face alignment for both steps. Results show more coherent SR faces than [BK00]. Similarly in [GBA+03], authors study two approaches when using super-resolution for Eigenface-based face recognition:

either computing a SR face from multiple LR observations and then refining it using HR Eigenface space, or using a linear mapping between LR features and the the Eigenface domain, that can then be used directly for recognition or reconstruction. This approach is extended in [PL08] with a local model of facial components and an extended morphable face model. Instead of PCA, in [YTMH08] the authors associate a global model using Non-Negative Matrix Factorization (NMF) (proposed in [LS01]) and a local enhancement using sparse coding, as proposed in [YWMH08]. In [ZFC+15], the authors address face hallucination using a bichannel CNN. The used data contains several degradations (Gaussian blur, motion blur) and resolutions. Constrained by the presence of face, the network learns a 2000 element dictionary to produce 100×100 HR faces using a first channel of fully connected layers. The resulting high-resolution components are fused with the upscaled input image, predicted with a second channel.

Another recent neural approach is proposed in [ZLLT16], where authors take profit of a cascaded architecture and high-resolution priors. To address faces "n the wild", *i.e.* not aligned faces, the procedure alternates between a dense mapping of facial high frequency priors and a high-frequency prediction step, which is added to the previous low-resolution input. For each step of the cascade, a network is learned.

Some global approaches such as those based on NMF [YTMH08] account for the part based nature of facial images which are composed of additive and locally independent parts (hair, eyes, nose, mouth, etc.). Many methods [LSZ01, LSF07, YTMH08, PL08] also use a local step to enhance the quality of the image. In the next section, Local approaches are presented, in which this aspect is explicitly exploited.

3.4.2.2 Local approaches

While global approaches aim to project the whole face into useful representation, local approaches focus on specific parts of the face. Such techniques may still require an *implicit* face alignment. For instance, if fixed position are considered, the faces must have been aligned as well. However, the different parts are not jointly projected into the same space, but rather processed individually.

A first approach that falls in this group can be found in [CZ01]. The authors use a MAP approach to fuse LR images with two different priors, based on spatial partitioning of the face into 6 subspaces. On each subspace, a PCA is performed on the training set. The FS-MAP (for Face Space-MAP) model gives SR face parts that belong to the computed PCA subspaces (as the optimization is performed over the PCA coefficients), and therefore present artefacts at the limit of two subspaces, such as between the nose and the eyes subspaces. The second relaxed IS-MAP (for Image Space-MAP) model

rather forces the SR face parts to live near their respective PCA subspace – and thus benefit from the learned representation of HR sub parts – but are more likely and artefact-free from a global perspective.

Using on fixed positions, the method presented in [JG08] is based on a tensor representation of a training database under four dimensions: people, modality (orientations), block (spatial region of the face) and resolution (pyramidal). Their approaches can therefore handle different orientations while performing SR. They first synthesise multiple modalities from a single input image, and then perform a local high-frequency enhancement. They also propose an iterative alignment that increase the quality of the SR faces. The same idea is exploited in [MZQ10], using fixed positions.

To handle the facial expression change of subjects in videos, the authors of [YB08] proposed a coarse-to-fine lattice-based alignment followed by a local fusion method. The deformation model aligns a given LR frame (*e.g.* containing open mouth and closed eyes) with a reference LR one (*e.g.* containing close mouth and open eyes). They show that the employed multi-frame methods benefit from this and allows to produce well-shaped HR images.

Finally, a neighbour embedding approach is proposed in [JHH+13]. The authors propose to refine a SR face by iteratively searching for its nearest neighbour on the HR manifold, using a constraint on the position of the processed patches *i.e.* considering local "nose" or "eye" manifolds. Later in [JHWH14], they extend the method by automatically learning an intermediate dictionary to have a more precise mapping.

Adaptive and relaxed local approaches In [CS14], the authors propose to use a facial keypoint feature detector to locate and improve the semantic components (eyes, eyebrows, nose and mouth). They first improve the resolution using an off-the-shelf SR algorithm. Then, using the detected facial components, a nearest neighbour search is done in a collection of high-resolution facial components examples. They fuse the chosen facial components with the image from the first step and produce high-resolution image. They propose to use different collection depending on the orientation of the face, as the appearance of the components change with the orientation of the face.

An extension for video can be found in [CCvBS15] with a consistency check on temporal domain

3.5 Conclusion

This literature review demonstrate the variety of methods that can be employed to perform SR. When a single LR image is available, example-based SISR approaches that automatically learn the relationship between LR and HR images from example pairs demonstrate good results. In particular, recent advances on Neural Networks make those approaches suitable for such task. They can be trained on large amount of data and process images in a single forward pass that does not require online search (*e.g.* for a nearest neighbour) or optimisation processes. However, little work has been made to combine the power of example-based framework and neural networks to address SR of specific image types such as text or face images, and evaluate the impact on the recognition engines that process them.

Contributions of this thesis on text single image SR The existing approaches in example-based text single image SR are promising, for both image reconstruction and improvement of automatic text recognition engines. However, there is still an important gap between the SR images and the HR ones, in both aspects. In this thesis, deep neural networks are used to produce more accurate SR images in chapter 4, providing an efficient way to learn specific models for text images. To foster research on such approach in the context of multimedia indexing, a database of text images extracted from televisual streams is presented, and the results of the first international competition on SR are reported and described.

Contributions of this thesis on face single image SR In a multimedia indexing context, faces are likely to appear with different poses, expression and illumination conditions. Therefore, approaches offering maximum flexibility would be favoured. In chapter 5, we introduce a two-step approach using neural networks. Inspired from [CS14], the proposed approach addresses face SR for lower resolution images and produces facial image closer to the original HR faces, and more easily recognisable by an automatic face recognition engine.

Contributions of this thesis on blind and robust SR The third contribution in chapter 6 will focus on making neural-based SR system robust to the variety of LR images, such as those obtained via different devices or under varying conditions. A very specific literature review is proposed at the beginning of this chapter, and a blind approach is described, that allows to produce accurate SR images without knowledge of the observation model.

Part II

Contributions

Chapter 4

Text Single Image Super-Resolution

Contents

4.1 Introduction

At Orange Labs, the MAS (Multimedia content Analysis and technologieS) team have been addressing the problem of audiovisual content indexation over the last ten years. Suppose you are given a collection of raw digital files (audio, video, image), without any indication on their nature or their content. Building services around those data like browsing, clustering or recommendation requires to extract the semantic content from them. Those semantic contents are generally keywords or ids, that allow to explore the media in a supervised (known classes and categories) or unsupervised (keywords without constrain and unconstrained discovery) way.

Though impressive performance are nowadays reached by automatic systems in Automatic Speech Recognition (ASR), facial recognition and object recognition, it decreases when such systems are applied to a degraded signal, pushing the robustness limit. In particular for visual content (image and videos), the resolution of the images is of great importance. Figure 4.1 depicts the performance of an OCR system on the *ULR-TextSISR-2013a* [NCGKO14] dataset for different resolutions. To address this, two strategies may be considered. The first consists in learning to recognize low-resolution objects. The second is to increase the resolution of the image. Those two approaches are "dual", and both include a kind of low-resolution image perception. In this thesis, we investigate around the second option, that provides a clear framework providing visual enhancement that can be "plugged" before any pre-existing automatic recognition engine. We shall consider the recognition engines as fixed black boxes, and study the evolution of their performance with the image resolution that the proposed methods allow to produce.

The first type of visual objects we address in this work is text. Natural images and frames extracted from TV streams often exhibit textual information. A very illustrative case is TV news, where text can indicate information about the speaker, the program, places, hour, etc. While useful for the watcher, those informations can also be automatically analysed using and Optical Character Recognition (OCR) engine to extract keywords associated with the image of the video. This has been addressed in previous work ([SKHS98, Lie03, EGS11, Ela13]). However, the OCR engines generally make the assumption that the detected text has sufficient resolution. This hypothesis can turn wrong in many cases as mentioned before: camera shake, distance to the sensor, low transmission rate. In those cases, text can still be detected by automatic text detection methods as they are generally more robust and detect LR texts in images as they exhibit relatively close properties or textures with HR texts. The character recognition, however, is more difficult as it requires well shaped individual letters.

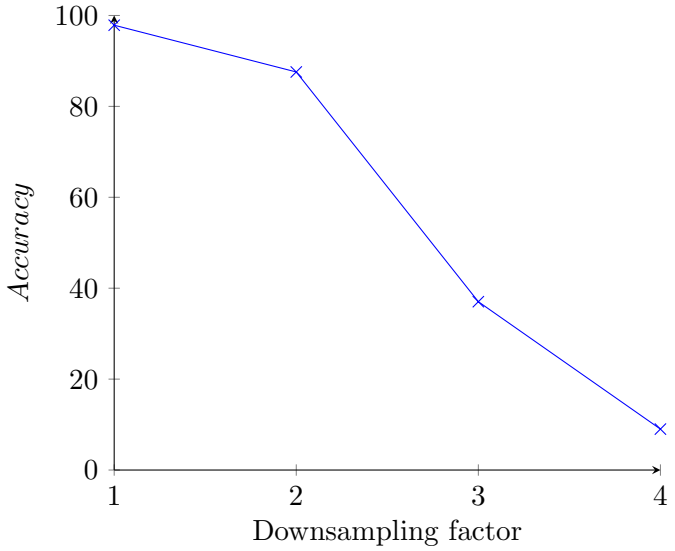


FIGURE 4.1: Variation of the Tesseract OCR performance on the *ULR-TextSISR-2013a* [NCGKO14] dataset with various text resolution. The horizontal axis represents factors of downsampling, compared with the high-resolution text. The texts are resampled at the HR sampling rate with bicubic interpolation.

The rest of this chapter is organised as follows. The postulate is presented in section 4.2: how adapting the data and providing a suitable learning framework allows to perform domain-specific SR. In section 4.3, we present the core of our approach. The results on the *ULR-TextSISR-2013a* [NCGKO14] document text image database are reported and analysed in 4.4. They indicate a clear improvement over previous sparse coding approaches. To address televisual contents containing text, a dataset was created and a competition organised for the ICDAR2015 conference. Those works are reported in section 4.5 and the results are analysed. Finally, a discussion on the specificity of the learned models is proposed in 4.6. As in other machine learning based application, there is a trade-off between the performance on a type of text images and the generalisation to new ones. Conclusive remarks are made in 4.7.

4.2 Domain-Specific SR using Data adaptation

As seen in the literature review (see chapter 3), learning-based methods are able to generate high-resolution images by capturing the relationship between the LR and HR spaces during the optimisation process. This may be done by learning the relationship of two manifolds (see paragraph 3.3.3.1), sparse dictionary learning (see paragraph 3.3.3.2) or neural networks (see paragraph 3.3.3.4). Although using natural image provides reasonable approximation of high-resolution images, we are now addressing a domain-specific SR problem that allows to make the assumption that it has inherent characteristics that have to be taken into account when proposing a solution.

As an illustration of this simple assumption, Figure 4.2 presents several histograms of difference images obtained by subtracting to high-resolution images their low-resolution and upsampled counterparts. The statistical difference between a natural image dataset (blue curve) and a text dataset (red curve) can be observed, for 8-bit integer images in the $[0, 255]$ interval. Text images, due to the presence of stronger gradients, exhibit more frequent high pixel-wise errors than natural ones, for which smoothness priors are well suited.

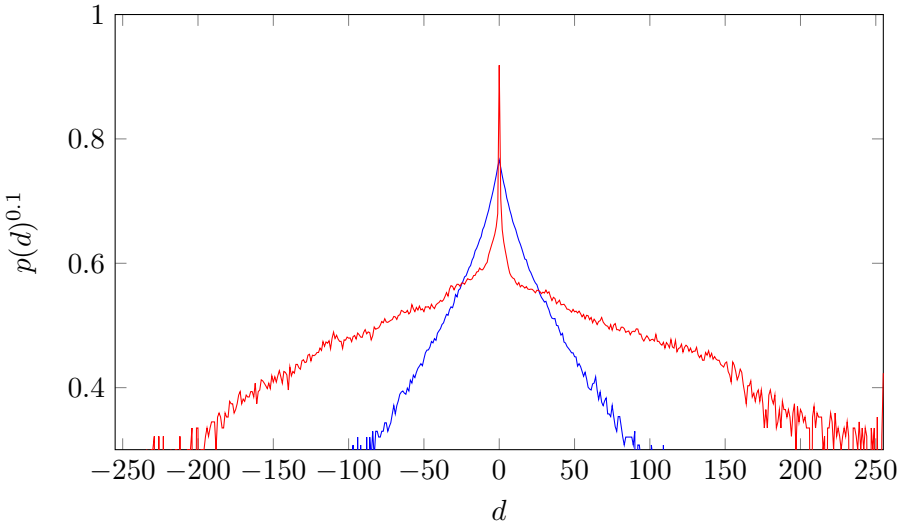


FIGURE 4.2: Errors histograms between the HR and interpolation images for natural (in blue) and textual images (in red), illustrating the different nature of data that may benefit from data adaptation during learning. The d horizontal axis corresponds to the pixel-wise intensity difference between high-resolution and interpolated images, and the y axis corresponds to the rate of occurrence (we plot $\sqrt[10]{p(d)}$ for a more comprehensive visualization).

To address the specificity of text images, several strategies exist. A first approach is to use domain-based knowledge to define heuristics that will help in this particular context. For text images, several assumption may be suitable. Some approaches capitalize on the fact that several occurrences of the same letter can be found in a given image, and consider this as a particular case of multiple image super-resolution at character level [LP07].

The data-based approach rather consider that the statistical properties of text can be captured by a learning algorithm if a suitable and adequate dataset is provided. Specific knowledge can be incorporated like in [WDL+13] where the dataset is composed of likely text strokes and therefore closed to the first mentioned approach, and is used to reconstruct all sort of text like ancient glyphs.

The approaches presented in the next section fall in this last example-based category. Example pairs of LR and HR patches are extracted from textual images, and form a training set with which a neural network can be trained.

4.3 Proposed methods for text image Super-Resolution

4.3.1 Method 1: Super-Resolution via Multi-Layer Perceptron

The first method we propose is to use multilayer perceptrons to perform super-resolution. Although MLP with one hidden layer were used in the past [Pla99, AGK95], we propose to investigate more advanced and deep architectures, by providing an extended study on the number of layers and neurons per layer. As seen in the definitions (see subsection 2.2.2 of chapter 2), simplest interpolation approaches consist in a fixed linear combination of the input pixel values. However, we can construct more powerful system that incorporate higher level of knowledge, using the edge information as described in section 3.3.1 of chapter 3. Statistically, most of the lost information in an interpolated image is located around edges, and flat regions are reasonably well reconstructed using basic interpolation techniques. Figure 4.3 shows an example of the difference image, obtained by subtracting an interpolated LR text image to the original HR one), *i.e.* the information that an ideal SR system would reconstruct.

FIGURE 4.3: Difference image (black for minimum negative values, white for maximum positive values) obtained by subtracting an interpolated LR text image to the original HR one.

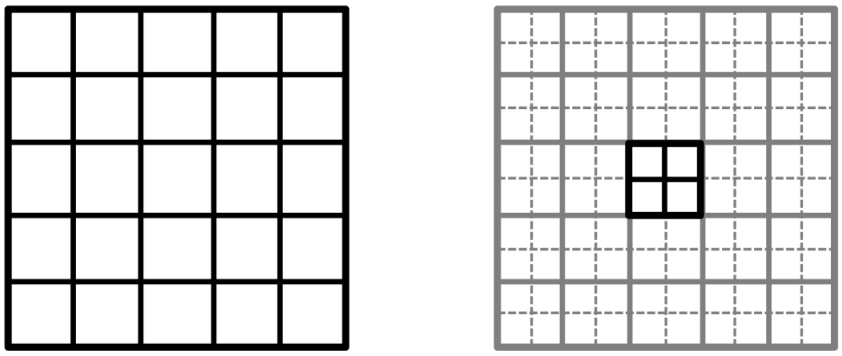
4.3.1.1 Designing the MLP

Multi-Layer Perceptron (MLP) is a feed-forward neural model that maps the input values to successive non-linear spaces, represented as neuron activations, until it reaches an output layer from which useful values can be extracted. The parameters of each layer are optimised to minimise a cost function over a training dataset. The design of the network can influence this minimization: along with the number of neurons (the system parameters), the depth of the network also plays a role. In the case of SR, the addressed inverse problem is not well-conditioned. Non-linear techniques such as sparse coding also gave significantly good results, in which the size of the overcomplete sparse representation is essential. To circumvent the observed limitations of linear regression

and shallow networks referred in paragraph 3.3.3.4 of chapter 3, we propose to use larger neural networks, with 2 hidden layers. For a better insight of the role of the number of neurons per layer, we will also evaluate single-hidden-layered nets.

4.3.1.2 Data representation and formatting

To train the neural network, we adopt a patch-based representation. This is a common data representation for example-based methods as indicated in the literature review (see subsection 3.3.3 of chapter 3). We propose to extract low-resolution patches directly from the available LR image, and predict high-resolution pixels corresponding to the central LR pixel. This is a similar set-up to [Pla99, AGK95]. For a scale factor of s, we aim to predict $s \times s$ pixels on the high-resolution grid from $M \times M$ pixels, as represented on Figure 4.4.



(A) $M \times M$ LR patch ($M = 5$) (B) $s \times s$ HR target pixels (upscaling factor $s = 2$).

FIGURE 4.4: The proposed method aims to analyse a low-resolution $M \times M$ patch, and predict the $s \times s$ high-resolution pixels, that are aligned with the central pixel of the low-resolution patch.

By doing this, we take the *a priori* that learning only those central pixels permits to efficiently focus on the most relevant area of the HR image. This differs from many methods of the literature that predict a large HR patch and perform a coherency check on the border. Predicting such a border is of course possible (as done in most of the sparse coding approaches reviewed in paragraph 3.3.3.2 of chapter 3) but would dedicate a part of the network parameters to this task. This is not optimal as more weights would be needed in the neural network, to connect the supplementary dimensions of the output targets, and to account for the increased complexity of the prediction. Averaging borders could also lead to more blurry images – although more complex fusion scheme could be used at the price of another increase in complexity.

Input patches normalisation We normalise the input patches by subtracting the central pixel value. This shares the same philosophy as LBP and is coherent with the fact that we predict the central missing HR pixels, which share the same zone of interest. While it also removes one dimension (the central LR pixel of the input patch always equals zero), we keep it for sake of representation simplicity.

Target patches normalisation The low-frequency content of the high-resolution image is highly correlated with the frequency content of the LR image. Even if dependent on the LR image formation model (see subsection 2.2.3 of chapter 2), the spectrum can be decomposed into one low-frequency and one high-frequency components:

$$I_{HR} = I_{LF} + I_{HF} \tag{4.1}$$

$$= (\uparrow I_{LR}) + I_{HF} \tag{4.2}$$

where $(\uparrow I_{LR})$ is the LR image resampled via interpolation on the HR grid, and I_{HF} is the residual high frequency that shall be predicted by the network. To avoid propagating the low-frequency information throughout the network, we propose to target only the high-frequency as in previous works in Sparse Coding, Regression or Neural Networks [AGK95, YWHM10, TDG13].

Scaling Although neural networks can use any real valued data, it is preferable to scale the data so that it shares the same scale as the internal representation used in the network. For instance, using a *tanh* activation function give neurons values in the $[-1.0, 1.0]$ interval, and data normalised to fit this scale is better, to avoid saturation of the activation functions in the network with small derivatives.

As the central pixel is removed, the pixel values can range from -255 to 255. We scale the input to ensure that values lie in the $[-1.0, 1.0]$ interval by dividing the normalised patches by 255. For the target, we allow larger values to be linearly inferred by an output layer by normalising by a value of two times the variance of the target values, computed from their histogram.

4.3.1.3 Architecture Selection

In order to have insights on how to design our MLP, we propose to test different architectures. The input layer of the MLP is the image patch. Each pixel is a node and will be connected to the first hidden layer of the network. The output layer is linearly connected to the last hidden layer of the network (*i.e.* no activation function is applied

to the output values). Various non-linear spaces dimensions (number of neurons per layer N) are studied, with one or two hidden layers (higher number of layers did not bring better performance). The tested architectures are presented in Table 4.1, where M^2 is the number of input pixels. To avoid ending-up with very high number of weights

TABLE 4.1: MLP architecture selection criteria

Architecture	**1st layer**	**2nd layer**	**Number of parameters**
MLP $(L=1)$	N_1	–	$\left(M^2+1\right) \times N_1 + (N_1+1) \times s^2$
MLP $(L=2)$	N_1	N_2	$\left(M^2+1\right) \times N_1 + (N_1+1) \times N_2 + (N_2+1) \times s^2$

in the proposed architectures, we limit the depth of the network to two hidden layers. The experimental results on document images are presented in section 4.4.

4.3.1.4 Model Optimisation

For each proposed architecture, we shall train a MLP using a training set and use a validation set to monitor the evolution of the learning process. We wish to optimise the network using stochastic gradient descent (SGD) so that it minimises an Euclidean loss function over the training set:

$$\Theta_{optimal} = \underset{\Theta}{\operatorname{argmin}} \sum_{i \in K_{train}} \| y_i - f_\Theta \left(x_i \right) \|_2^2 \tag{4.3}$$

where Θ is the set of weights (or parameters) of the MLP, y_i is the ith normalised target patch from the K_{train} training samples, x_i is the ith normalised input patch and f_Θ is the network's non-linear function.

In order to adjust the weights, we will start with a random initialization of the network's weights using a zero centered Gaussian with a small variance. A stochastic gradient descent using standard backpropagation algorithm will then be performed using a fixed learning rate λ.

4.3.2 Method 2 : Super-Resolution via Convolutional Neural network

The second method involves using Convolutional Neural Networks (CNN) instead of MLP. We shall keep the same patch representation and normalisation. MLP do not consider the $2D$ nature of the input patch and connect each of its pixels to all the neuron of the first layer. This results in a large number of weights (or parameters) in the network. CNN allow to reduce this number of parameters by applying the same

learned convolution filters to the different positions of the input patch. The result of each convolution corresponds to a feature map, that can be further processed by other layers. Moreover, the CNN keep track of the $2D$ nature of the patches throughout the successive non-linear filters, while the MLP do not have internal $2D$ representations.

4.3.2.1 CNN design for SR

Convolutional Nets allow to extract a spatially structured representation. At each level of representation (each layer), local features are expected to be found independently from their exact location. The resulting feature maps are usually spatially reduced via pooling (subsampling) layers as the presence of a feature is more important than its exact location. In our case, where we focus on very low-level characteristics of the LR image signal, we do not use pooling layers. It would make the final prediction harder to localise, which is moreover meant to lie on a denser spatial grid. However, this ability to cascade filters is very interesting as we seek for a non-linear transformation from the input low-resolution data to the high-resolution data. In the MLP method proposed in subsection 4.3.1, we saw that the whole spatial information has to be extracted in the first layer, as it is the only one that has access to the raw $2D$ data. With a convolutional network, we can keep track of the 2D nature of the data on a deeper scale, along the layers. Under this relaxation, we can expect more interesting low-level features to be captured by a CNN, as non-linear spatial features can be extracted by upper layers as well.

4.3.2.2 Architecture Selection

Again, we propose to evaluate the impact of the architecture design on the SR performance and on the complexity of the model. We still consider the output pixels to be linearly predicted from the same space. In practice, the CNN is designed to extract features in a convolutional manner and map them to the high-resolution space. To simplify the architecture selection, we adopt a non-fully connected scheme between the first and the second layer, similar to those proposed in [GD04]. This scheme integrates fosters two behaviours:

- **Specialization:** each convolution map is connected to two kernels of the second layer, that learn convolution kernels that are specific to this kind of map.
- **Fusion:** every map of the first layer is connected pair-wise to kernels, that learn convolutional kernels

This sparse hand-crafted connection scheme allows to reduce the number of parameters. Given a number of maps in the first layer of N_{C1}, we end up with N_{C2} maps in the second layer:

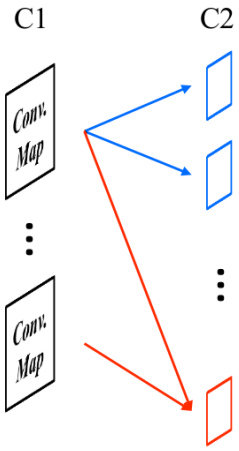
$$\begin{aligned} N_{C2} &= (2 \times N_{C1}) + \left(\frac{N_{C1}!}{(N_{C1}-2)!2!} \right) \\ &= N_{C1} \frac{(N_{C1}+3)}{2} \end{aligned} \tag{4.4}$$

for which we hold Θ_{C2} parameters (including bias):

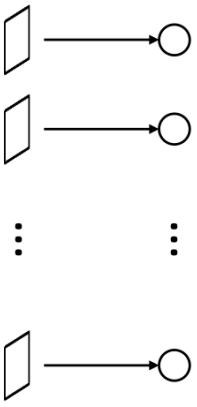
$$\Theta_{C2} = N_{C1} \left(2(1 + f_{C2}^2) + (1 + 2f_{C2}^2) \frac{(N_{C1}+3)}{2} \right) \tag{4.5}$$

where f_{C2}^2 is the filter's size.

This is illustrated in Figure 4.5 – (A). The convolutional feature maps are followed by



(A) The convolutional layers with a hand-crafted sparse connectivity. In blue, connections fostering specialisation; in red, connections fostering fusion.



(B) Neuron layer connected to the convolution maps, in 1-to-1 fashion.

FIGURE 4.5: Connection used between layers in the CNN architecture.

a neuron layer. This can be done in a 1-to-1 fashion (see Figure 4.5 – (B)), where the neurons learn spatial feature for each map independently, or in a fully connected fashion, where all maps are fed in each neuron with a large increase in complexity. We evaluate the impact of the architecture in the next section, along with the first method using MLP.

4.4 Application to document image SR

To evaluate the proposed approach on meaningful data, a dataset proposed for document image SR is used. This allows to evaluate and compare the performance for image reconstruction (how close the SR image is from the original HR image), but also for text recognition (how recognisable the reconstructed text is). Both aspects are studied as well as the learned neural networks to gain insights on the best architectures.

4.4.1 The ULR-TextSISR-2013a dataset

This dataset was presented in [NCGKO14], where the authors proposes a selective patch processing scheme when using patch-based approaches. They report reconstruction and OCR mesurements for several learning-based approaches which can serve as a comparative basis.

Data description The dataset is composed of black text over a white background, similar to the content present in document images (see Figure 4.6). It contains distinct training and testing data. The training data is composed of a single line containing all characters present in the set. The testing data is composed of 5 paragraphs from the Peter Pan book, containing $13,428$ characters. Each paragraph is generated using three different font families and two different font sizes, yielding six different images per paragraph. In addition, the text is composed of a mix of normal, italic and bold font styles that bring more variety and challenging tasks.

Synthesis According to [NCGKO14], while the test images were generated from a pdf, the training images are generated using imagemagick [1] for both glyph and image rendering.

Downsampling For downsampling, the Matlab *imresize* function is used, with $'bicubic'$ option, which corresponds to a bicubic interpolative kernel with antialiasing.

4.4.2 Experimental set-up

Based on the *ULR-TextSISR-2013a* dataset, we propose convenient changes to the training data and present the evaluation scheme, similar to the one proposed for the original dataset.

[1] http://www.imagemagick.org

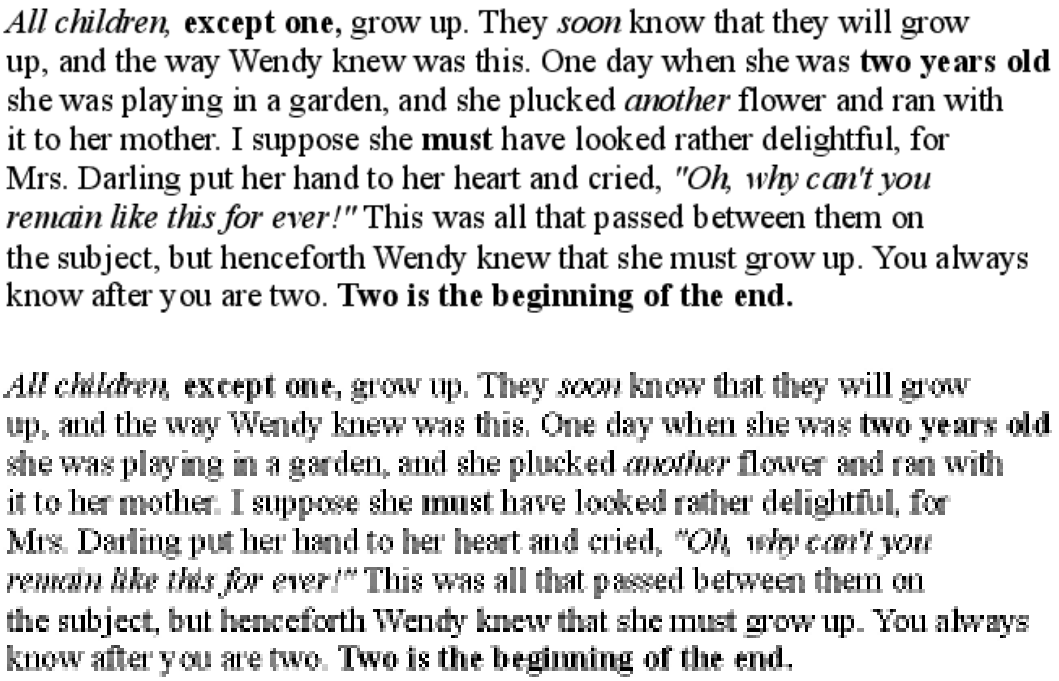


FIGURE 4.6: Example of a LR and HR image pair extracted from the *ULR-textsisr-2013a* dataset [NCGKO14].

4.4.2.1 Training data

In order to have a richer training dataset for our system, we propose to produce training images using the GIMP [2] software, and using the same procedure as described in [NCGKO14]. We extract a text from wikipedia and generate it accordingly to the original training data.

As proposed, we extract pairs of patches from the low and the high resolution images. We randomly extract the pairs (1,000 pairs per image) with a non-zero requirement for the targets. For a patch to be kept, the absolute sum of the normalised target patch must be superior to $\ell = 0.01$, which accounts for at least one pixel that is non-zero. This ensures that flat patches are not over-represented in the constructed training set.

4.4.2.2 Evaluation measurement

We use the same measures as [NCGKO14] to evaluate the performances of the different proposed methods: Mean Squared Error (MSE), Peak Signal to Noise Ration (PSNR) and OCR accuracy.

[2] https://www.gimp.org/

1. MSE reflects the squared difference in gray levels between two images A and B of dimensions X by Y.

$$MSE = \frac{1}{X \times Y} \sum_{x \in [0,X]} \sum_{y \in [0,Y]} \Big(A(x,y) - B(x,y) \Big)^2$$

2. Employed in signal processing, PNSR gives a more absolute meaning to the reconstruction, given the maximum value the signal can reach. It is still closely related to the MSE.

$$PSNR = 10 \times \log \left(\frac{255^2}{MSE} \right)$$

3. The Structural Similarity differs from the first two measures by taking into account the structure around each pixel (in a sliding window) and gives a better indication on the visual quality of the resulting image.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_x^2 + C_2)}$$

   The overall score is obtained by averaging the SSIM at each position in the image, also known as MSSIM.

4. When processing text images we can produce a joint evaluation of both standard measures and classification, recognition or detection scores. Optical Character Recognition systems allow to produce an accuracy measure for evaluation. The Character Error Rate (CRR) is used, which is the Levenstein distance between recognized characters and ground truth transcription, divided by the total number of characters. Following the proposal of [NCGKO14], the performance of the proposed approaches are evaluated using the same tools (Tesseract OCR 3.02 and UNLV-ISRI accuracy tool).

For a fair comparison, and even if not made explicit in the original paper, the reported CRR results do not take into account ground truth spacing characters, although including the related errors. This means that the accuracy score is calculated using:

$$CRR = \frac{N_{Err,chars} + N_{Err,spacing\ chars}}{N_{chars}} \tag{4.6}$$

The dataset contains $N = 13,428$ characters, of which $N_{spacing\ chars} = 2,550$ are spacing characters. Note that the authors only take $N_{spacing\ chars} = 2,258$ spacing characters.

4.4.3 Results and analysis

4.4.3.1 Quantitative Results

We use the same measurements as in [NCGKO14]: image reconstruction measurements are accomplished via MSE and PSNR, while the OCR improvement is characterized in term of CRR.

In Table 4.2, we report the results obtained with a single-hidden layer MLP.

TABLE 4.2: Performances of a 1 hidden-layered MLP on the *ULR-textsisr-2013a* test set, with increasing number of neurons N_1.

Config.	1-a	1-b	1-c	1-d	1-e	1-f
N1	10	25	50	100	200	500
Complex.	544	1,354	2,704	5,404	10,804	27,004
PSNR	21.61	22.24	22.59	22.88	**22.89**	22.68
MSE	21.53	19.99	19.20	18.57	**18.54**	19.00
SSIM	0.947	0.957	0.959	0.960	**0.961**	0.956
CRR	89.57	92.20	92.09	92.70	92.53	**93.03**

From this first experiment, we can see that increasing the number of neurons from the hidden layer does improve the quality of the predicted SR image. However, the results seem to saturate between 100 and 500 neurons, where the same order of performance is obtained. However, we can see that the best accuracy is not reached for the best performing neural network in term of reconstruction. We comment on this result in paragraph 4.4.3.2.

In the second experiment which results are reported in Table 4.3, we have two hidden layers in the MLP. We start from a low number of neurons to compare with the previous experiment, and choose to increase it progressively in both layers.

We observe a similar behaviour to Table 4.2 with an asymptotic level of performances. We can already notice that adding a second layer allows to break the asymptotic limit observed with only one layer. This pleads in favour of deeper networks, and not only more neurons per layer. This is also corroborated by the fact that deeper 2-hidden-layered configurations obtain systematically better reconstruction results with an equivalent complexity, *i.e.* with the same number of degree of freedom. This is the case for configurations $\{1-a/2-a\}$, $\{1-b/2-b\}$, $\{1-e/2-e\}$.

TABLE 4.3: Performances of a 2-layered MLP on the *ULR-textsisr-2013a* test set, with increasing number of neurons in the two hidden layers.

Config.	2-a	2-b	2-c	2-d	2-e	2-f	2-g
N1-N2	10-10	10-50	50-50	50-100	50-150	100-150	100-200
Complex.	654	1,254	5,254	8,004	10,754	20,754	26,004
PSNR	22.01	22.67	23.05	23.25	23.63	**24.15**	24.05
MSE	20.52	19.03	18.20	17.80	17.05	**16.03**	16.23
SSIM	0.9478	0.9601	0.9625	0.9649	0.9695	**0.9748**	0.9727
CRR	91.97	93.28	92.85	93.73	93.82	**94.69**	94.44

The third experiment concerns the proposed CNN configuration. Recall that it holds 3 non-linear layers, 2 of which are convolutional with a sparse connectivity and "one-to-one" connected layers. This allows to limit the number of weights while having a deeper network, and keeping the spatial convolution practical aspect that allows to generate directly the output image. These results in Table 4.4 show the pertinence of neural

TABLE 4.4: Performances of different ConvNet configurations on the *ULR-textsisr-2013a* test set.

Config.	3-a	3-b	3-c	3-d	3-e	3-f	3-g	3-h	3-i	3-j
C1	2	4	8	12	16	20	24	28	32	40
Complex.	185	498	1,520	3,070	5,148	7,754	10,888	14,550	18,740	28,704
PSNR (dB)	20.49	21.59	22.89	23.25	23.88	23.79	24.18	24.16	**24.55**	24.48
MSE	24.36	21.51	18.48	17.74	16.47	16.67	15.91	15.95	**15.27**	15.39
SSIM	0.918	0.946	0.956	0.958	0.962	0.964	0.968	0.969	0.963	**0.972**
CRR (%)	90.35	90.70	93.35	95.08	95.23	95.10	95.49	96.09	**96.42**	96.13

networks for the SR regression task. As shown in Figure 4.7, all the configuration reach better reconstruction than bicubic interpolation and the reported results of [NCGKO14]. Moreover, configurations using deeper networks (MLP with two hidden layers or CNN) lead to improved recognition.

The performance observed with the CNN configurations is superior to the one observed in the MLP. First, the efficiency is increased as we obtain better results in both reconstruction and OCR accuracy for the same number of parameters (see $\{2-f/3-i\}$ for instance). Second, we obtain unreached results in term of accuracy and reconstruction.

Training and Convergence We chose high learning rates without observing divergent behaviour, and lowering it did not result in better convergence. This is mainly

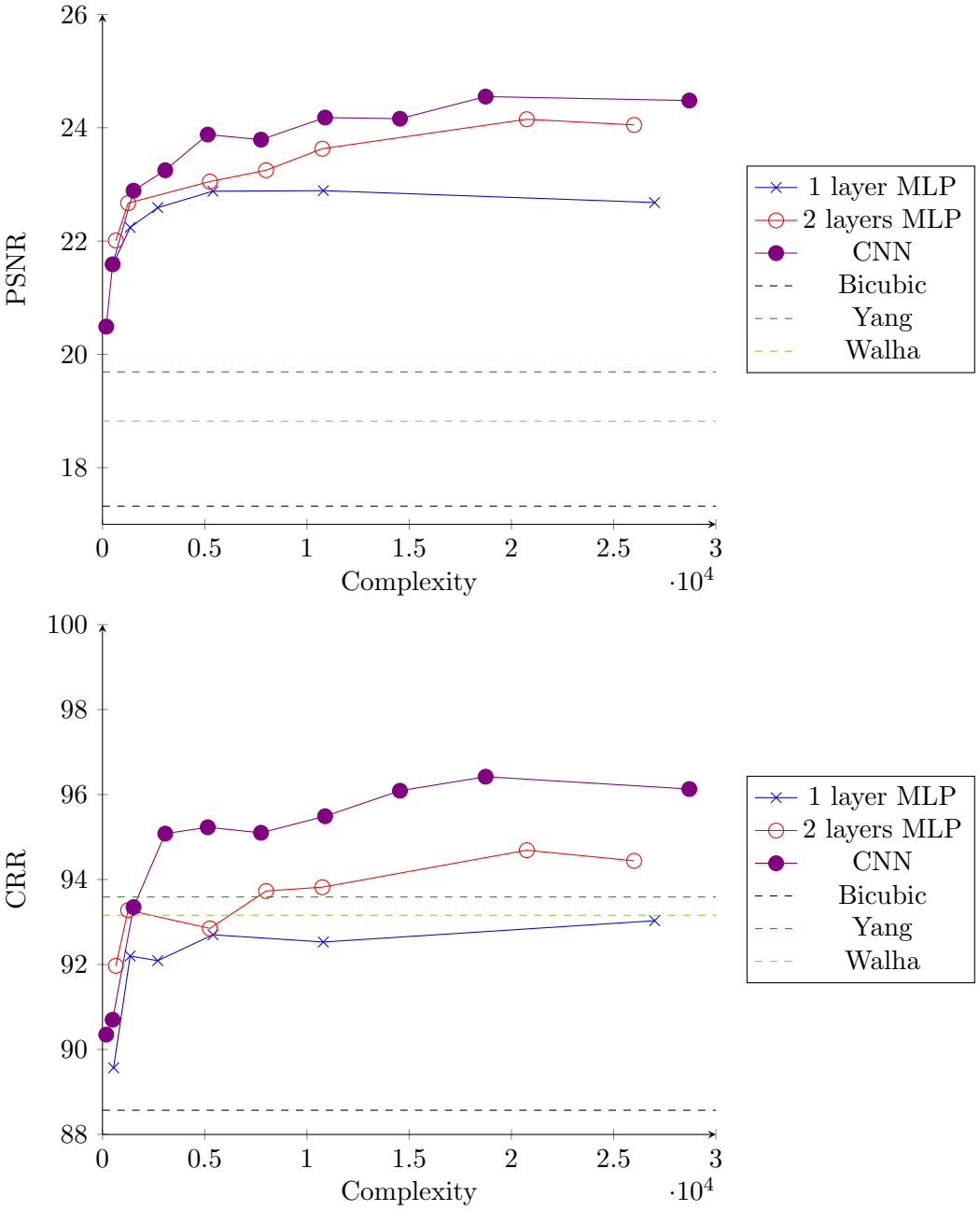
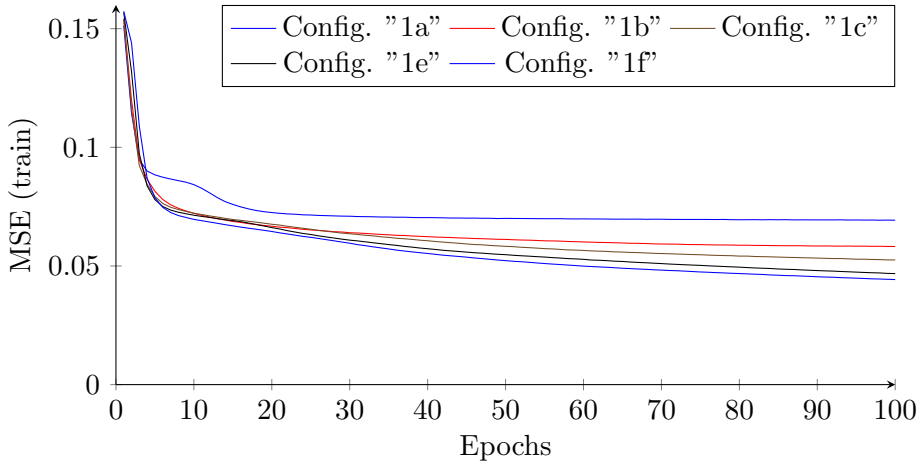


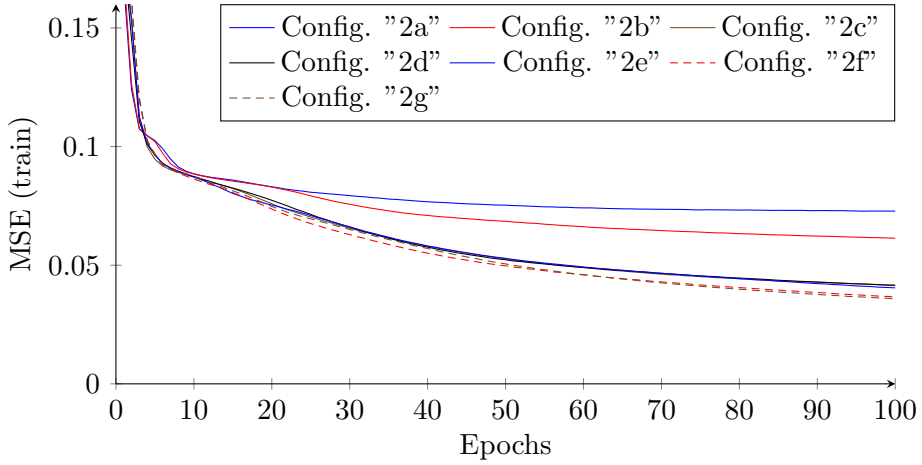
FIGURE 4.7: Obtained results for the various tested architectures. The depth of the networks matters as the MLP with two hidden layers outperforms the shallow one with a single hidden layer, for equivalent number of parameters. The use of CNN further improve the results. The previous results obtained in [NCGKO14] are outperformed by the deep networks.

due to the architecture of the network and the adaptive learning rates depending on the number of units. We also used low momentum values (0.2) that allow more dynamics in the learning procedure from a sample to another.

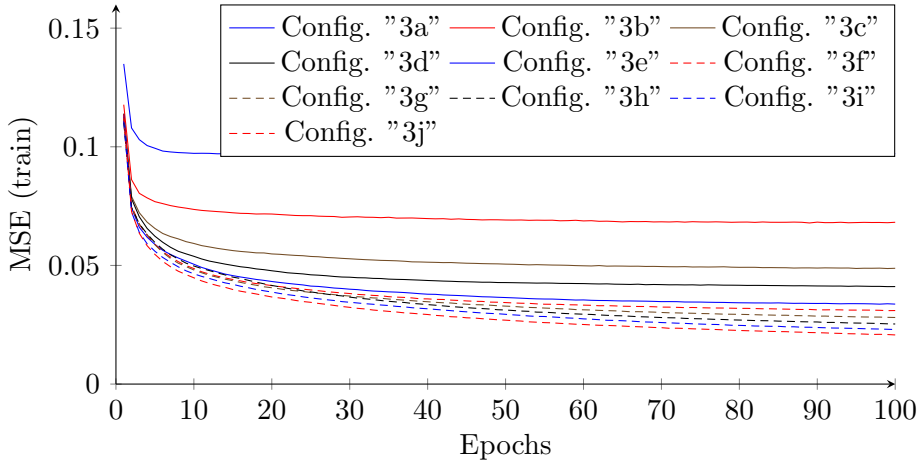
Figure 4.8 represents the evolution of the objective function on the training data for the different models, which is adequate with the obtained score.



(A) Mean Squarred Error loss evolution during training, for MLP architectures with one hidden layer.



(B) Mean Squarred Error loss evolution during training, for MLP architectures with two hidden layer.



(C) Mean Squarred Error loss evolution during training, for CNN architectures.

FIGURE 4.8: Evolution of the cost function during the training. Deep architectures (B and C) allow to decrease the cost function compared with the shallow one (A). The learning also benefits from an increase in the number of parameters for each architecture category.

4.4.3.2 Qualitative and OCR-based evaluation of the obtained SR images

To better understand the relationship between the pixel-level differences measures by the classical PSNR/MSE/SSIM, we can take a closer look at the obtained SR images. Moreover, even if there is a trend of obtaining better OCR accuracy when better reconstruction is reached, this is not always true, as shown in Table 4.2 where the best OCR accuracy is reached for the third-best reconstruction performance (PSNR/MSE). We propose a deeper analysis in the next paragraphs.

Qualitative & visual analysis We can already focus on the visual reconstruction of the SR images and compare them with the LR and the ideal HR ones. Figures 4.9 and 4.10 show those differences for two cases that constitute the most obvious characteristics that are seen with bare eyes.

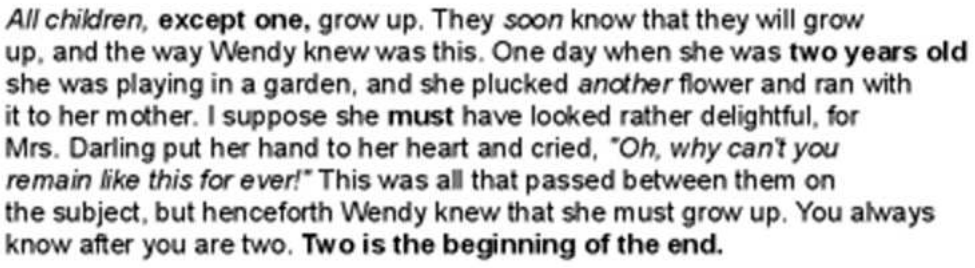
The cases of rare letter association is interesting. For complex situation, the network seems to give an average answer, as depicted for the "ok" transition in word "looked", printed in Figure 4.11.

A third interesting case is the font family which is not present in the training set: we notice different behaviour. A first result is that simple structures are still well reconstructed, and the blurry effects are removed. At this scale, the result can be expected as edges are almost similar. Another noticeable behaviour is the "hallucination" of letters, such as the "e" letter in Figure 4.12.

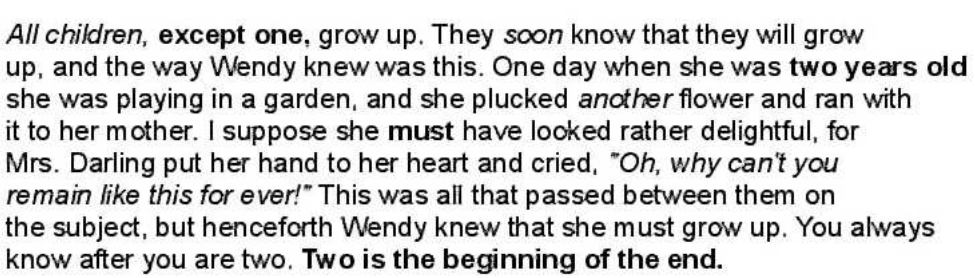
We also observe noise around letters. This is due to a response of the network to the stimuli contained in the retina, although the central pixel would not need correction. Along with that, a "phantom" noise sometimes appears in the white background, where a tiny residual value is added to the white images. The visual analysis can be enriched by the per-letter accuracy that shows where the different models take their respective advantages. Note that the spacing characters can account in the final result, because a missed spacing character is still added to the error counter.

OCR performance analysis Throughout this study, we used the *accsum* tool[3] that gives very precise reports and statistics about the errors between the groundtruth and the obtained OCR results. They include analysis by class of character: `Spacing Characters`,`Special Symbols`,`Digits`,`Uppercase Letters`,`Lowercase Letters`, but also a detailed representation of the largest edit distances used to compute the final score.

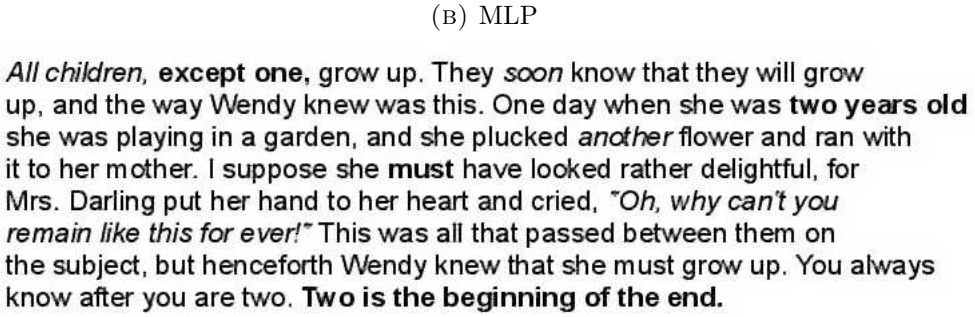
[3]Originally developed by UNLV/ISRI [RN96] and available at https://github.com/eddieantonio/isri-ocr-evaluation-tools



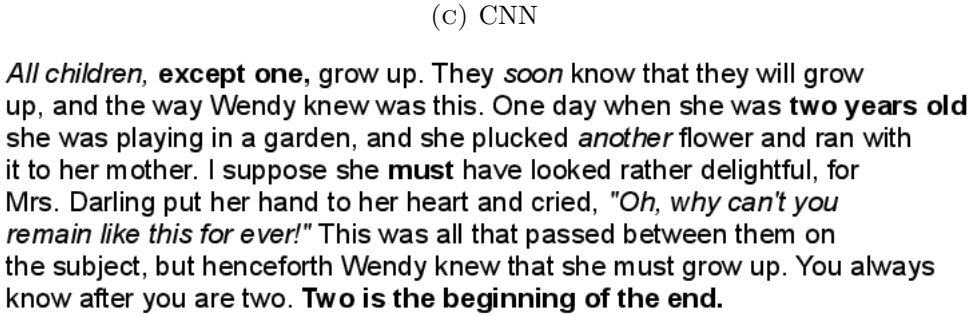
(A) Bicubic



(B) MLP

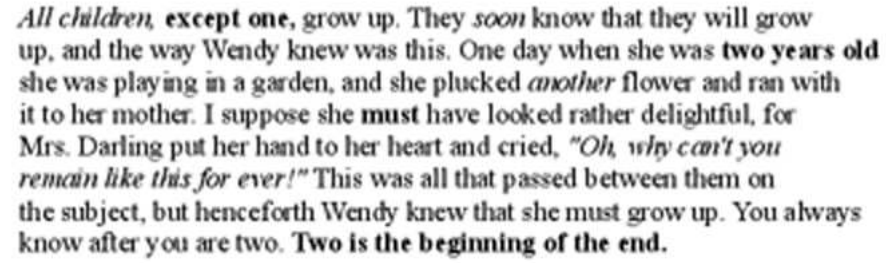


(C) CNN

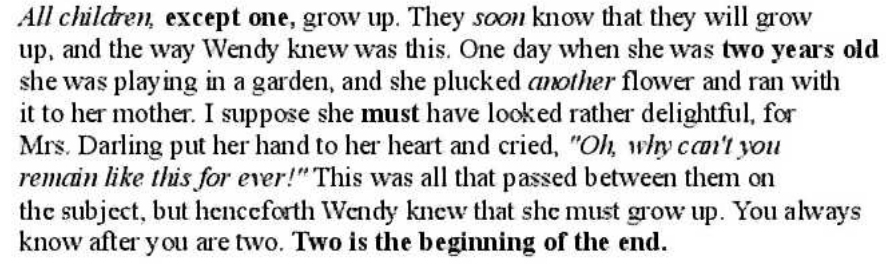


(D) HR

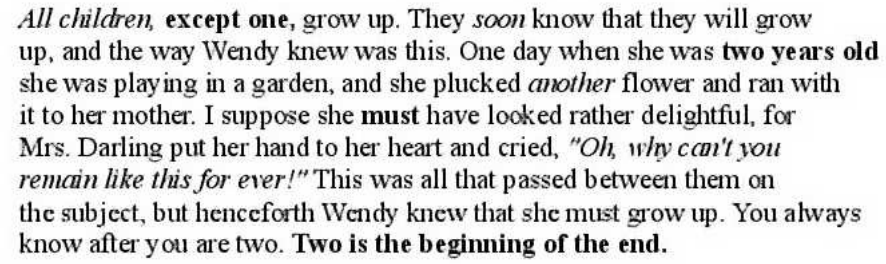
FIGURE 4.9: Results for the Arial, 10pt text for Bicubic interpolation, MLP, CNN. The proposed SR mehods allow to reduce blur artefact and ambiguous patterns such as inter character spaces or fine dots.



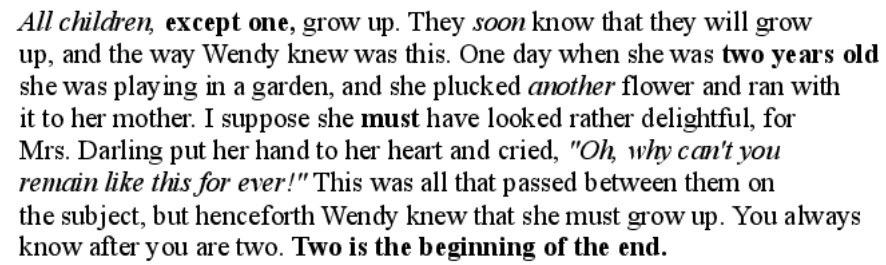
(A) Bicubic



(B) MLP



(C) CNN

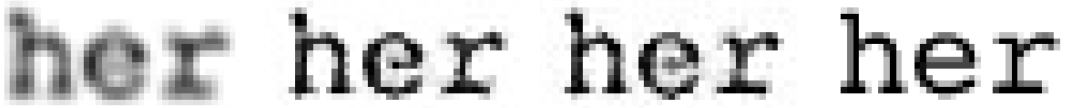


(D) HR

FIGURE 4.10: Results for the Times, 10pt font for Bicubic interpolation, MLP, CNN. More artefacts are noticeable as the font is more complex (mixed low-resolution strokes, serif).

(A) Bicubic (B) CNN

FIGURE 4.11: A complex case where the transition is not well corrected: the strokes of the "k" letter are not well reconstructed while similar to the "l", and the "o" seems deformed by the presence of the complex structure of its ambiguous neighbour.



(A) Bicubic (B) MLP (C) CNN (D) HR

FIGURE 4.12: For the Courier font that is not present in the training dataset, some strokes such as vertical and horizontal edges are well reconstructed, while others seem "hallucinated" in different ways. Here, the "e" letter is well predicted from the MLP model while poorly inferred by the CNN mode, which seems to draw ambiguous pixels instead of a straight horizontal stroke.

As mentioned in paragraph 4.4.2.2, we adopted the evaluation protocol of [NCGKO14], where we do not take into account white-spacing character. This is a pertinent choice as they tend to be well classified by the OCR even using a simple interpolation.

Table 4.5 shows the overall performance per class. We propose to focus on different points:

- **Successfully reconstructed letters** *i.e.* the letters that benefit most from super-resolution, compared with interpolation
- **Increased error rates** the "hardest" letters to be reconstructed
- **Surpassing the high-resolution images** – when the SR images give better results than the HR ones.

Punctuation marks As many marks occur for only few pixels in the low-resolution image, they can end up being ambiguous after SR. Specifically, comas are the marks that suffer most from the low-resolution, and seem to be often classified as points by the OCR. Both MLP and CNN based SR systems seem to reconstruct better those characters, and CNN is especially good for commas (-14 errors / $+\,7.78\%$). However, none of them achieve to get closer to the high-resolution performances (only five missed comma by the OCR on the HR data).

TABLE 4.5: Comparative analysis of the OCR results obtained with the best MLP and CNN configurations, with bicubic and groundtruth, high-resolution.

		Bicubic		MLP (config 2-f)		CNN (config 3-i)		HR	
Count	Character	Missed	Right	Missed	Right	Missed	Right	Missed	Right
6	!	4	33.33	2	66.67	1	83.33	1	83.33
12	”	6	50	1	91.67	1	91.67	0	100
12	’	5	58.33	2	83.33	3	75	0	100
180	,	66	63.33	63	65	49	72.78	5	97.22
6	-	1	83.33	0	100	0	100	0	100
150	.	30	80	9	94	10	93.33	6	96
6	1	1	83.33	2	66.67	0	100	0	100
6	4	1	83.33	2	66.67	0	100	0	100
6	:	0	100	0	100	0	100	0	100
12	;	6	50	4	66.67	3	75	1	91.67
6	A	0	100	0	100	0	100	0	100
6	B	0	100	0	100	0	100	0	100
36	D	0	100	0	100	0	100	0	100
6	E	2	66.67	0	100	0	100	0	100
24	H	0	100	0	100	0	100	0	100
12	I	0	100	0	100	0	100	0	100
36	M	5	86.11	3	91.67	4	88.89	0	100
6	N	0	100	0	100	0	100	0	100
24	O	8	66.67	3	87.5	3	87.5	2	91.67
12	S	1	91.67	0	100	0	100	0	100
30	T	1	96.67	0	100	0	100	0	100
36	W	10	72.22	3	91.67	1	97.22	1	97.22
6	Y	0	100	0	100	0	100	0	100
6	[	0	100	0	100	0	100	0	100
6	]	1	83.33	0	100	0	100	0	100
738	a	48	93.5	18	97.56	12	98.37	6	99.19
144	b	1	99.31	5	96.53	0	100	1	99.31
234	c	22	90.6	9	96.15	6	97.44	0	100
396	d	20	94.95	26	93.43	10	97.47	3	99.24
1338	e	68	94.92	27	97.98	1	99.93	0	100
162	f	33	79.63	9	94.44	13	91.98	6	96.3
222	g	4	98.2	2	99.1	0	100	0	100
744	h	44	94.09	28	96.24	20	97.31	0	100
528	i	77	85.42	15	97.16	19	96.4	5	99.05
6	j	0	100	0	100	0	100	0	100
132	k	1	99.24	4	96.97	6	95.45	4	96.97
354	l	31	91.24	8	97.74	2	99.44	0	100
252	m	17	93.25	5	98.02	1	99.6	0	100
738	n	48	93.5	41	94.44	12	98.37	13	98.24
792	o	41	94.82	50	93.69	14	98.23	37	95.33
204	p	6	97.06	1	99.51	0	100	0	100
6	q	0	100	0	100	0	100	0	100
606	r	58	90.43	18	97.03	16	97.36	6	99.01
660	s	40	93.94	24	96.36	21	96.82	3	99.55
882	t	99	88.78	55	93.76	32	96.37	6	99.32
336	u	18	94.64	13	96.13	9	97.32	2	99.4
90	v	11	87.78	4	95.56	2	97.78	0	100
372	w	24	93.55	16	95.7	7	98.12	12	96.77
36	x	1	97.22	0	100	0	100	0	100
246	y	16	93.5	15	93.9	13	94.72	1	99.59
12	z	2	83.33	4	66.67	3	75	2	83.33

Successfully reconstructed letters We propose to focus on letters for which the amount of missed (misrecognised) characters drops by half compared with bicubic interpolation. With that criterion, we can note that for both systems, the largest benefits goes to the letters `(E,W,a,c, e, f, g, i, l, m, p, r, v)` for the MLP, and letters `(E, W, a, c, d, e, f, h, i, l, m, n, o, p, r, t, u, v, w)` for the CNN, highlighting the higher performances of the CNN models.

Increased error rates Contrary to the previous observations, some letters do not benefit from the increase in resolution and high-frequencies provided by the SR models. On such letters, the bicubic interpolation obtains better per-letter recognition scores. This is noticeable for letters `(k, o, z)`.

Surpassing the high-resolution images Occasionally, we can observe the case where the Super-Resolution outperforms the HR images. This is the case for letters `(b, w, n)` where the CNN has higher recognition score. This is of course not generalised in our case. Moreover, as for the previous remarks, the internal dynamics of the OCR can also play a role (language model), and it may be dangerous to extrapolate on the statistics of isolated letters. However, this confirms the interest of learning-based super-resolution for recognition task and indicates that the learning-based approach may overtake the groundtruth results and benefit from the large amount of examples it has seen to propose effective solutions.

4.4.3.3 Analysis of the learned networks

The power of neural approaches is to learn and perform end-to-end tasks. In SR, hand crafted features are often used and sometimes associated with dimensionality reduction methods, before being mapped to the HR space. Here, we simultaneously learn feature extractors (first layers of neural connections), the mapping and the reconstruction of the HR high-frequency. Although understanding the contribution of each component of a deep neural network is tricky and a research subject in itself, we can look at some parts to make reasonable assumptions. Notably, we can examine the learned filters, closest from the input data. We can also analyse the low-level contribution of some cells in the network.

Learned filters Observing the learned filters, we can notice that some of them behave like simple feature extractor (derivative filters $a9, c9$, with blue borders in Figure 4.13), close from classical hand-crafted ones and produce densely activated maps. For instance,

the four filters ($a1-4$, with green border in Figure 4.13) with higher variance for the MLP are close to diagonal edges extractor , while others such ($a10, b1, b3, b7$ with orange border in Figure 4.13) have vertical or horizontal aspects.

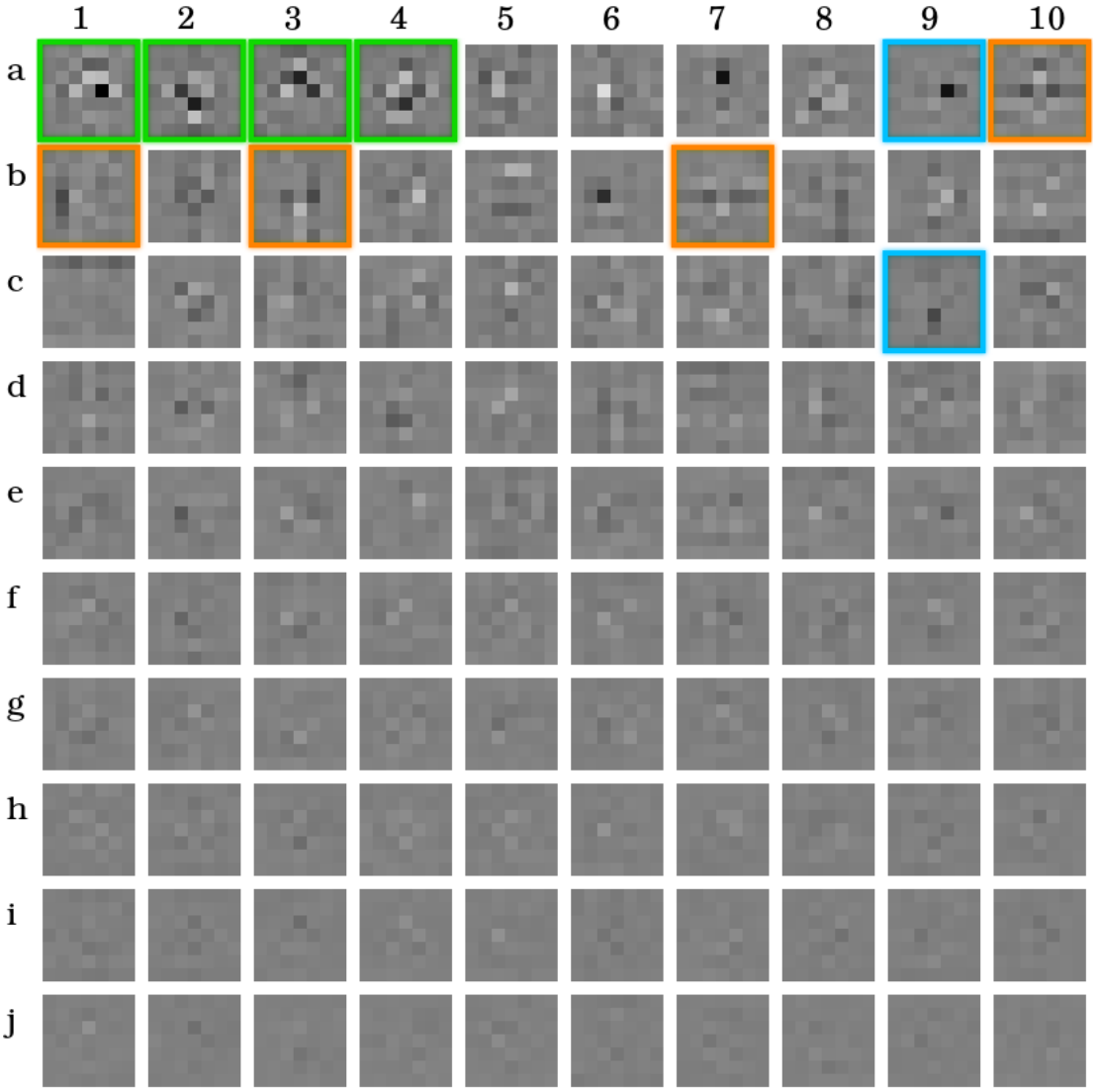


FIGURE 4.13: 9×9 weights learned by the first layer of the best MLP architecture, ordered by variance. Some of them exhibit comprehensive aspects, such as vertical (orange) and diagonal (green) edge extractors, or derivative filters (blue).

For the CNN, the filters are smaller as the patches are spatially processed by different layers. While several filters with very low variance appear to be learned by the MLP, the CNN filters result in simpler ones and more balanced variance. The CNN relaxes the capture of complex spatial patterns as the forward layers can associate the non-linear response of the first layer.

Apart from the mentioned filters, we can see that some others are less obvious to explain. However, as the performance of the network increase with the number of filters, we can

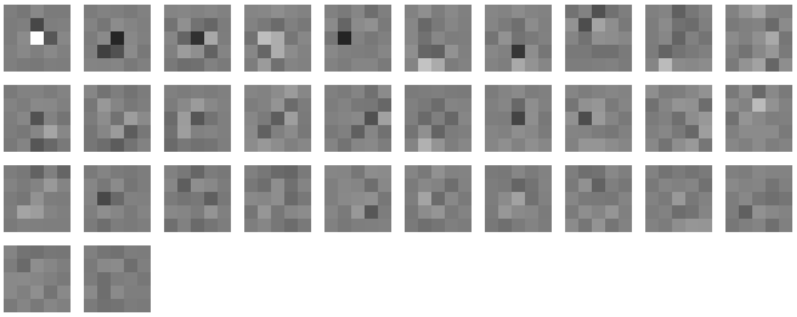


FIGURE 4.14: 5×5 filters of the first convolutional layer learned by the best CNN architecture; ordered by variance.

deduce that the network relies on such filters, even if they have less interpretable role. In the next paragraph, we visualise the latent maps to gain insights on the internal processing of the neural networks.

4.4.3.4 Optimisation of the proposed architecture

Normalisation scheme simplification for faster convolutional processing The normalisation scheme chosen for this experiment is patch-based, and is non-trivial to perform on a full image. The usual schemes (local or global) perform a normalisation on the whole image before treating it via a convolutional neural network. In [DLHT14, DLHT16], the authors do not perform any normalisation of the input and the output. Here, keeping the benefits of the residual targets at output, we evaluate the performance of the network on non-normalised input. The obtained results are reported in Figure 4.19, and show similar performance. The training time is also similar and indicate that the convolutive nature of the first layer can get rid of the steady low-frequency contained in the input patch. While the initial normalisation is patch-based (*i.e.* each patch is processed independently, with overlap), this optimisation allows to visualise the full activation maps on a given image, and have some visual clues on the behaviour of the network. The first and second layers appear to act as dense non-linear feature extractors. While the first layer sticks with interpretable features, it is much harder to predict how the informations may benefit to the final prediction. The last non-linear layer contains more "simple" activations maps. Each map appear to have a spatial role and the activation function seems to saturate each position in most cases. The contribution of each map to the final output (*i.e.* the learned weights of the last layer) is displayed in Figure 4.17, and can be seen as overcomplete subpixel basis that are summed up according to the activated maps to form the output image.

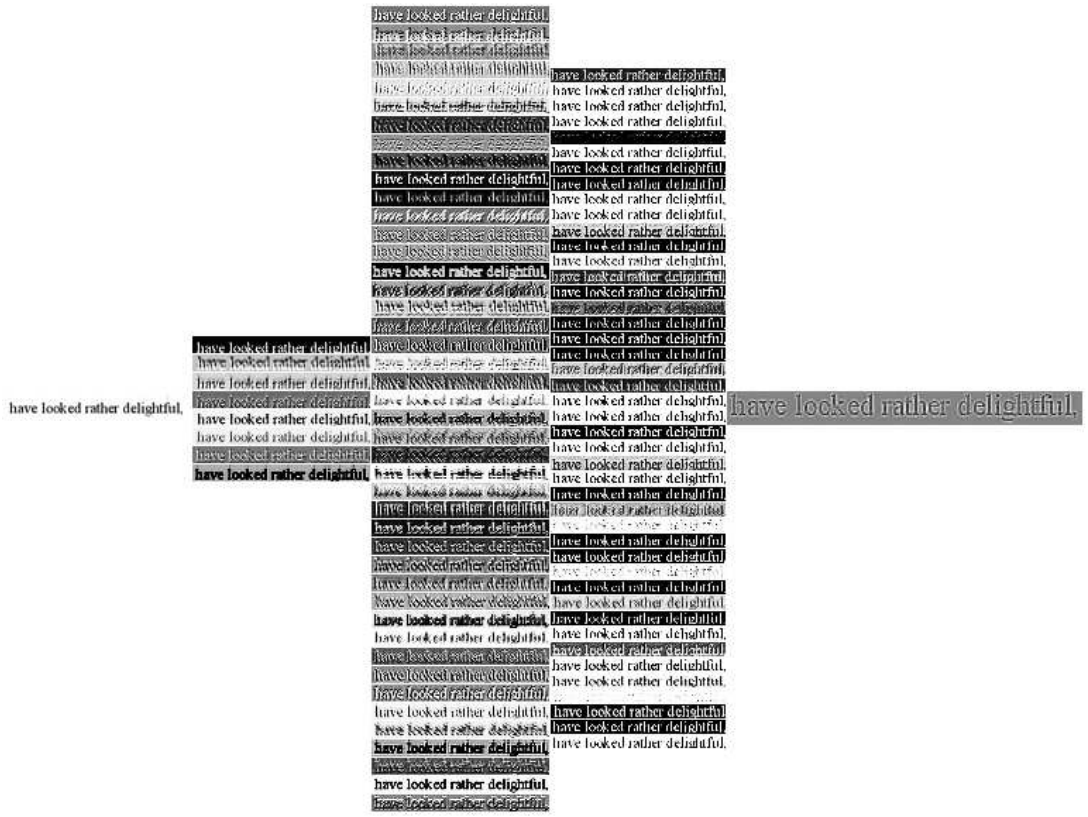


FIGURE 4.15: Internal activations of the CNN for a text image. While the latent maps have complex appearance corresponding to the spatial neural activation, the last layer exhibits the attributes of the HR-bicubic difference, accordingly to the training objective.

(A) Activation maps in the first layer

(B) Some activation maps in the second layer

FIGURE 4.16: Activation maps from the first (A) and the second (B) layers of the network, for an input image composed of a line of text. The first layer maps (A) have interpretable appearances as they are similar to high-pass filtered images. The second layer maps exhibit more complex spatial behaviour as they are a non-linear combination of the first ones.

Spatially, most of the maps are densely activated. However, some of them have a

FIGURE 4.17: Basis weights (learned) of the linear output layer. Each 2×2 output patch is a linear combination of these basis, weighted by their respective neuron activation.

more sparse or specialised behaviour, only on specific position such as stroke ends (Figure 4.18a), or small horizontal strokes (Figure 4.18b).

(A) Example of maps activated on stroke ends.

(B) Example of maps activated on small edge strokes.

FIGURE 4.18: Various observations in the final layer activation maps.

Architecture vs. Equivalence in the number of free parameters The best performing network is obtained for about $18k$ parameters ($18,740$ exactly). It is however difficult to evaluate the impact of the architectural variations on the performances of the system. To evaluate this, we use a similar architecture as configuration 3-c of Table 4.4 but with a fully connected layer as the third layer (instead of one-to-one connections). This gives the same order of parameter ($18,548$).

On Figure 4.19, we observe a slight drop in accuracy, but within the same order of accuracy for OCR performance (-0.5 points, 95.96 instead of 96.42). Additionally, due to a slight correction in the example selection, the PSNR score is superior as no "phantom noise" is present ($25.88dB$, versus $24.55dB$ for the best configuration 3-i in Table 4.4).

What is also lost in this configuration is a high dimensional space before the linear output layer. We see a trade between compactness configuration and speed. However,

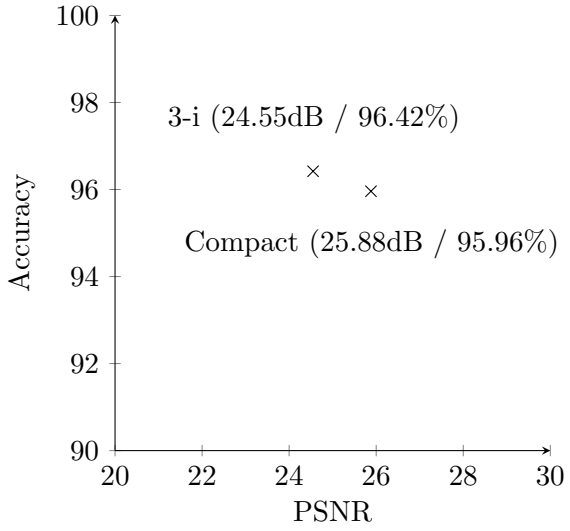


FIGURE 4.19: Results obtained with a more compact architecture, with equivalent number of parameters. Similar performance is obtained for recognition, and a better PSNR is obtained with cleaner images due the absence of the observed phantom noise (see 4.4.3.2).

as many weights of the last layer were close to zero, a reduced space seems convenient. In the following, we shall use it for complementary experiments.

4.4.3.5 Complementary results

To analyse the compromise between the convergence speed (that depends on the quantity of data for a same number of training epoch over the whole training set) and the accuracy. As shown before, the training time is reasonable, using high learning rates (10^{-3}). In a first time, we focus on the impact of a rough quantity of data (and thus the diversity). In a second time, we analyse how the pruning strategy may be further improved over the simple scheme of removing non-flat targets.

Influence of the number of training samples Recall that we randomly extracted $120,000$ pairs of patches from the training images during the experiments. We now analyse how the number of training samples can impact the learning procedure. We shall reuse the configuration proposed in 4.4.3.4 as it is more compact. We propose to train randomly initialized network with 4 different numbers of patches: $1,2k$, $12k$, $120k$, $1,200k$. As depicted on Figure 4.20, training with more samples gives better results on the *ULR-textsisr-2013a* test set. This allows us to reach unprecedented scores in both accuracy and reconstruction measures for the highest number of samples (96.71% and 26.85 dB). However, lowering the number of samples decreases dramatically the performances, even if we observe rapidly an increase in the scores, compared with bicubic.

The training time is linearly increased with the number of training samples, *e.g.* takes 0.26 seconds per epoch for $1,200$ samples, *vs* 2 minutes and 20 seconds for $1,200k$ samples, on a i5 M520 CPU with 4 cores.

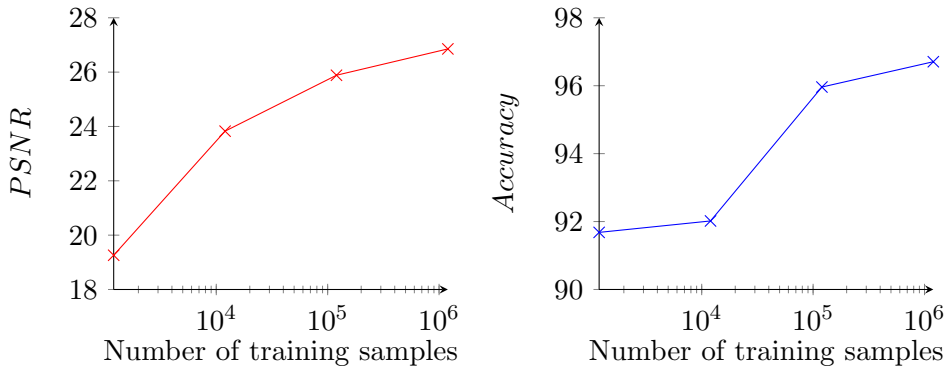


FIGURE 4.20: PSNR and OCR accuracy improvement over quantity of training data (plotted on a log-scale). The networks benefits from more training samples.

Complementary results using original training data A closer look into the *ULR-textsisr-2013a* dataset indicates that there is a lack of compliance between the generated images for training and testing. Indeed, it seems that the anti-aliasing process used for producing the high-resolution test images (from a PDF document) is different from the one used for producing the high-resolution training images (using imagemagick). We end up with images that are more similar to the testing images, even if differences are still observable : subpixel shifts are present and impact the anti-aliased rendering, letter spacing is less accurate than the original training data, and An example of the observed difference is displayed on Figure 4.21.

To see how this impact the results, we train the previous used architecture in similar conditions (120,000 training samples) on this training data. This also allows a complete and fair comparison with the results reported in [NCGKO14]. We observe a degradation of the results. Even though the observation model is the same, the difference in HR images produce more ambiguous strokes. As we observed that using ten times more data helped us to improve the score with the proposed training data, we also train a network with $1,200,000$ training samples. Using this "data trick", we get closer to the expected results (see Figure 4.22), and still achieve significant improvement over the results reported in [NCGKO14].

Apart from the high-resolution, the other factor that likely affects the results is the difference letter occurences. The provided training images contain all possible characters and pairs of characters. First, it is quite hard to gather all the strokes present in those images using a random patch selection. Second, these generated images do not reflect the statistical reality of the English language as the letters appear independently of their

(A) Bold Arial font letters in the original training set.

(B) Bold Arial font in the proposed training set.

(C) Bold Arial font in the original testing set.

FIGURE 4.21: Example of the differences between high-resolution images in the training and testing datasets, due to the underlying antialiasing process when synthesising the high-resolution letters. While still not exactly the same (different letter spacing, aliasing and subpixel shifts), the process we used produces similar images to the test images.

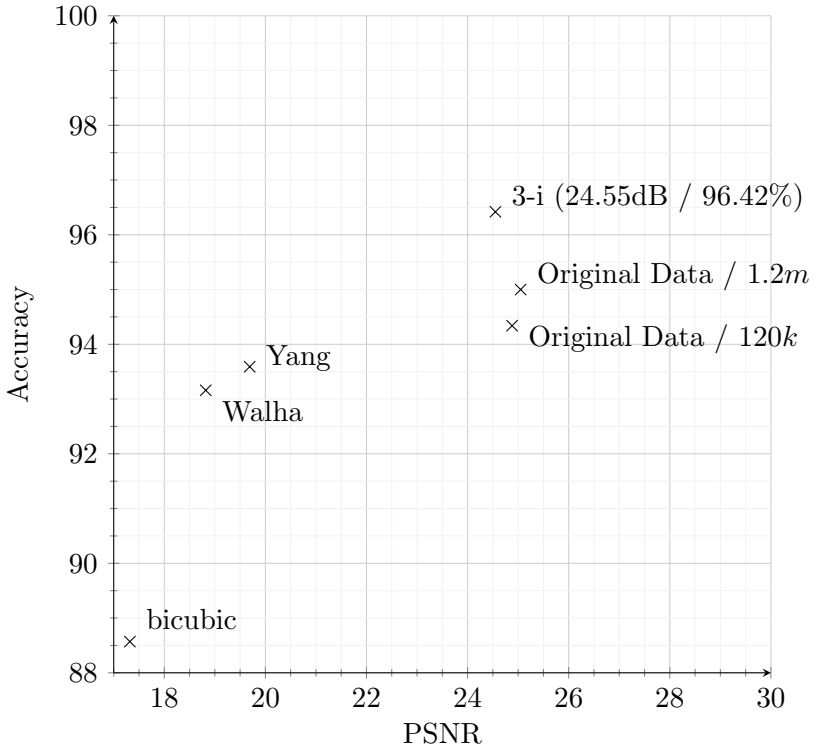


FIGURE 4.22: Results obtained with original data (using the compact architecture), we observe competitive results but a decrease in accuracy due to a less precise stroke reconstruction coming from the difference outlined in 4.4.3.5

actual appearance probability. Using english text probably serve us as the training text images share the same letter statistics with the test text images.

4.5 Super-resolution of TV-based textual content

The proposed approach have been so far tested on a very specific document image set, that do not reflect all the type of texts that can be found in other contexts. For multimedia indexation, text can be found with a lot more variability of fonts, colours, shapes, orientation than the one present in this first study. In this section, the creation of a dataset with text contained in TV streams is presented. The organisation and the results of the first international competition on text single image SR based on this dataset are reported and analysed, and compared with the approach proposed in section 4.4.

4.5.1 Motivation

Televisual contents are very rich and diverse, ranging from cinematographic productions to live news broadcasting. With the increasing quantity of programs and the new opportunity for broadcasting brought by the Internet, a need for indexing such content in order to be able to browse inside it has risen. Along with speech analysis and recognition, facial identification or clustering, detecting and recognizing text embedded in the video may be a precious source of information on the semantic content of the video. In news broadcasting, it may provide information about language, identity, geography, subjects, keywords or time.

However, in many situations such textual information may be of poor resolution. While HD streams are now common on broadcast television, they may contain many amateur videos taken with non-professional equipment (phones, webcam, hand-held devices) which have reduced resolution or quality. Such streams can also be shared or reproduced in reduced resolution over the Internet. Moreover, distant scene text can still be challenging to exploit even in from HR sources.

4.5.2 Creation of the *ICDAR2015-TextSR* dataset

The only dataset for single text image super-resolution was the one used in this first part of our work from section 4.3. In order to propose a new SR evaluation framework, and adequate images for the context of low-resolution televisual text content, we proposed a new publicly available database[4]. The main criteria we established were:

- Relevant nature of images, from real-world use case,

[4] https://liris.cnrs.fr/icdar-sr2015

- Challenging text size, fonts and backgrounds,
- Double evaluation scheme with available high-resolution images for reconstruction measures and text annotation for OCR accuracy performance evaluation.

The protocol is presented in the following paragraphs.

Video selection and text image extraction We extracted the text images from French TV HD video streams, in 5 different channels, on various types of TV shows (news reports, sport, investigation, entertainment, advertisement). The text was detected using a similar approach to [DG08] and a draft recognition was performed automatically using [EGMS14] to prepare annotation. We also added non-detected text manually and removed poor samples with poor quality. We cropped the text according to the bounding boxes generated by the detector, and adjusted it when necessary to fit integer multiples of the different downsampling factors. Examples of such obtained text images are shown in Figure 4.23.

FIGURE 4.23: Examples of cropped text from HD TV streams proposed in the *ICDAR2015-TextSR* dataset.

Low-resolution images generation To generate the low-resolution images, we chose to use Matlab's `imresize` function, with `bicubic` option, which applies an anti-aliased bicubic kernel for resampling factor inferior to 1.0. This method was found to be the

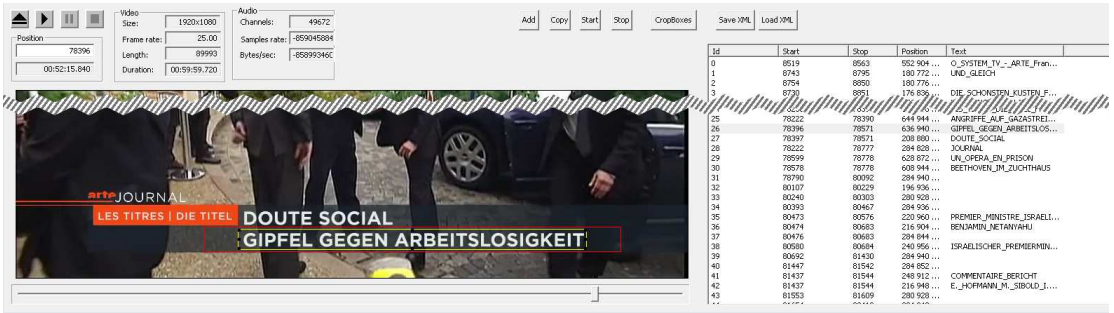


FIGURE 4.24: Annotation software developed at Orange Labs, used to annotate the dataset.

FIGURE 4.25: Different resolution (top: LR, middle: SD, bottom: HD) for three types of images. Left: a simple example with white text over a dark background, center: complex background, right: severely degraded image. For reading purpose, all images in this figure are upscaled to the same size. Better seen in digital form.

TABLE 4.6: Number of images and characters per set.

	Number of images	Number of characters
Train	567	12,565
Test	141	2,929
All	708	15,494

most representative to mimic the downsampling process in MPEG videos. We started from the HD resolution images (called HD images) and downsampled them by factors of 2 (SD images) and 4 (LR images). From a video standard perspective, SD images roughly correspond to SD resolution and LR images to CIF resolution.

Ground truth Two types of ground truth are available. The first one is the high resolution reference frames, which is the standard to evaluate SR methods: SR images are compared with the original ideal HR image, which gives a reconstruction error, that can be expressed in terms of pixel-wise or more advanced measures (see 4.5.3.1). The second one is text annotation: since images contain text, each of them was manually annotated, based on the draft automatic annotation (see 4.5.2).

Train and test The dataset is divided into two subsets: a training and a test set. Each of them was randomly sampled from the whole set, with a 80/20 ratio, resulting respectively in 567 and 141 images, for 12,565 and 2,929 characters (see table 4.6).

4.5.3 ICDAR2015 Competition on Text Image Super-Resolution

Competitions are common in many conferences, and allow to foster research on specific subject, share datasets and provide common ground for evaluation. In particular, the

International Conference on Document Analysis and Recognition (ICDAR) takes place every other year, and hosts many competitions since its creation (*e.g.* robust reading [KGBN+15]). However, no conference had seen a SR competition in the past, on any kind of images. In order to promote SR approaches to the document community and, the other way around, specific applications such as text image to the SR community, we proposed a competition based on the dataset described in the previous subsection 4.5.2.

For this first competition, participants were only required to produce $\times 2$ super-resolved images (SR images) from a set of LR images, to recover SD images. During the competition, participants had access to the fully annotated training set, and only to the low-resolution images of the test set, without annotation. The full dataset was made available after the competition, which contains all the images data (HD, SD and LR). The perspective is to provide data to address higher upscaling factors.

4.5.3.1 Evaluation procedure

As previously, the SR results are evaluated using two different kinds of measures: reconstruction measures and OCR accuracy score.

Reconstruction measures The PSNR, RMSE and MSSIM measures (see previous section, paragraph 4.4.2.2) were used to evaluate how close from the original images the SR ones are.

OCR accuracy score The Character Recognition Rate (CRR) was used to determine the OCR score of the different images, and evaluate the impact of the proposed SR approaches on the Tesseract OCR 3.02[5] system.

4.5.3.2 Competitors proposed methods and results

Seven teams registered for the competition. A total of four sets of results were received, from three different teams. A description is given for each of them.

ASRS - Wahla et al. The ASRS system [WDL+15] (Adaptive Sparse Representation Selection based system) was submitted by Rim Walha, Fadoua Drira, Franck Lebourgeois and Adel M. Alimi, as a result of a collaborative work between the REGIM laboratory (ENIS, Tunisia) and the LIRIS laboratory (INSA-Lyon, France). Sparse coding is the

[5] https://code.google.com/p/tesseract-ocr/

core technique of the ASRS system proposed to enhance the spatial resolution of textual images. The underlying idea of this technique is to represent an image patch as a sparse linear combination of elements from a suitably chosen dictionary. Motivated by the key role of the dictionary in sparse coding, the proposed resolution enhancement system is based on the use of multiple learned dictionaries and is adapted to the specificities of writing patterns. More precisely, it includes two phases: the learning phase and the reconstruction phase. The key idea of the first phase is to find more appropriate dictionaries adapted to the particular specificities of characters and learned from a well-clustered training LR/HR patch-pair database. To improve the unsupervised clustering of this database, an intelligent clustering method is applied and a new local feature descriptor, referred to as Histogram of Structure Tensors (HoST), is introduced making it possible to capture the local information of an image patch [WDL+15]. Via the proposed descriptor, the clustering performance is improved and the learning phase can provide more appropriate dictionaries representing each cluster. Given multiple learned dictionaries, a reconstruction phase is designed in order to adaptively select the appropriate dictionary that is useful for improved recovery of each local patch.

SRCNN - Dong et al. The SRCNN system [DLHT14] (Super-Resolution Convolutional Neural Network) was submitted by Chao Dong, Ximei Zhu, Yubin Deng, Chen Change Loy, and Yu Qiao from the Shenzhen Institutes of Advanced Technology Chinese Academy of Sciences, the Chinese University of Hong Kong (SIAT-CUHK). The method trains an end-to-end convolutional neural network, as described in [DLHT14]. It obtains state-of-the-art performances in natural images. The network takes the low-resolution image (after interpolation and padding) as the input and directly output the high-resolution one. According to the stochastic properties of text images, the authors propose a four-layer network and investigate different network designs (e.g., filter size). They also conduct model combination to further improve the performance.
The proposed network contains four convolutional layers. This i^{th} layer contains n_i filters of support $f_i \times f_i \times n_{(i-1)}$, where $i = 1, 2, 3, 4$ and $n_0 = 1$. The $max(0, x)$ function is chosen as the activation function in the first three layers. The basic parameter settings are $f_1 = 9$, $f_2 = 7$, $f_3 = 1$, $f_4 = 5$, $n_1 = 64$, $n_2 = 32$, $n_3 = 16$, denoted as $9 - 7 - 1 - 5$. Authors have also investigated structures $9 - 7 - 3 - 5$, $9 - 7 - 5 - 5$, $9 - 5 - 5 - 5$, $11 - 9 - 7 - 5$, $11 - 9 - 9 - 5$, $13 - 11 - 9 - 5$ and $15 - 13 - 11 - 5$. 30 image pairs are selected from the provided training set for validation, and the rest 537 image pairs for training. All low-resolution images are upscaled by a factor of 2 using bicubic interpolation in advance. 156,941 18x18 sub-images are cropped from the 537 image pairs as the training set. All networks are trained with 5,000 iterations.
Authors found that combining the outputs of different networks can largely improve the

performance. In total, they have successfully trained 11 networks of different structures or initialization parameters. They use a greedy search to find the best model combination. First, they find the model that achieves the highest PSNR on the evaluation set. Then, they combine its results with that of another model from 11 networks. The combination with highest PSNR is saved as the best 2-model combination. Similarly, they can identify the best 3 20-model combination. At last, they choose the best one as their final submission. Two results were submitted: the first one with the best OCR score, and the second one with the best reconstruction score, based on the validation subset. The same evaluation is conducted with the OCR score. This leads to 2 sets of results, one favouring the reconstruction score and a second one favouring the OCR accuracy score.

Synchromedia Lab - Farrahi et al. The proposed system [MC10] was submitted by Reza Farrahi Moghaddam and Mohamed Cheriet from the Synchromedia Lab, ETS, University of Quebec. The proposed super-resolution method for text images is a generalization of the in-house super-resolution method developed for natural images, as described in [MC10]. The method is built on top of three main components, among other ones. The first component is the grid-based multi-scale approach to modelling in development of a multi-scale binarization method. It is worth mentioning that the grid-based approach is general and can be used in any type of modellingg. This approach was used here in order to build a fast and multi-scale solution to super-resolution problem. The second component is inpainting of undecided pixels. Finally, the third component is a deblurring step using specially-designed point spread functions. In addition, other processes are considered in order to control and contain the level of blur even before the deblurring step. Furthermore, other binarization, segmentation, and text detection techniques are used in order to adapt the method to text and non-text regions. Training and model selection was performed for almost all the processes involved in the proposed method in order to adapt them to the dataset provided in the competition.

4.5.3.3 Baseline Methods

To provide a full comparison basis, we proposed some results using basic upscaling methods and publicly available approaches to SR among the best in the state-of-the-art, that take advantage of training data.

Interpolation Methods Upscaling or interpolation methods are just ways to add intermediate discrete values computed as a weighted combination of a neighbourhood, and do not add any prior knowledge necessary to overcome the SR ill-posed problem but

their simple, core equations. The baseline comprises two common interpolation methods that give the best reconstruction results: bicubic and Lanczos3 (see paragraph 2.2.2.2 in chapter 2).

State-of-the-art methods We also present results using Zeyde et al. [ZEP10] and A+ [TDSVG14] SR methods. We retrain the models using the public code and the training data provided to the competitors.

Our method We also presented, on an indicative basis, the obtained results using the approach proposed for document images in section 4.3. Similarly, we trained a CNN to map between LR patches and HR high-frequency information. We utilise a light version of the architectures described in paragraph 4.3.2.2, with $N_{C1} = 4$ maps in the first layer, $N_{C2} = 14$ maps in the second layer, and a fully connected layer with 14 neurons, yielding only $2,076$ parameters. 9×9 LR input patches are used to generate 2×2 output HR pixels. The full training set is used for train and validation.

4.5.3.4 Results of the competition

In this subsection, we present for each SR method the score obtained on the *ICDAR2015-TextSR* test set and give a short analysis.

Results for baseline, state-of-the-art, and submitted methods are reported in table 4.7. We also report results for our method which was not submitted to the challenge, as we organised it.

We can observe a general trend in favour of the submitted learning based approaches, that give the best overall performances. The SRCNN method gives the best results, in term of image reconstruction (set 2) and also OCR (set 1). Note that the original SD images have an OCR score of 78.80. However, each method can perform differently depending on the content of the LR images. Generally, to get a good and realistic reconstruction, using learning based method will yield good results as it learns the mapping between low and high resolution information. This competition being constrained to TV text images, this specific mapping will benefit from specific training data, as provided in this dataset. For OCR accuracy, the letters have to be well shaped. This of course happens with a good HR reconstruction, but even a method that do not perform well in terms of reconstruction can yield better results in terms of OCR accuracy. This is one advantage of the double evaluation. A typical example is the method proposed by [MC10]: while the reconstruction scores are lower in terms of PSNR, RMSE and SSIM

TABLE 4.7: Results of the described methods (baseline, ours, state-of-the-art, and submitted) on the *ICDAR2015-TextSR* dataset.

Method	Reconstruction measures			CRR
	RMSE	PSNR	MSSIM	
Bicubic	19.04	23.50	0.879	60.64
Lanczos3	16.97	24.65	0.902	64.36
Ours	11.27	28.25	0.953	74.12
Zeyde et al. [ZEP10]	13.05	27.21	0.941	69.72
A+ [TDSVG14]	10.03	29.50	0.966	73.10
Synchromedia Lab [MC10]	62.67	12.66	0.623	65.93
ASRS [WDA+14]	12.86	26.98	0.950	71.25
SRCNN-1 [DLHT14]	7.52	31.75	0.980	**77.19**
SRCNN-2	**7.24**	**31.99**	**0.981**	76.10
Original HR	–	–	–	78.80

than the standard bicubic or Lanczos3 one, their SR scheme (locally adaptive) can lead to better OCR score on some images (see table 4.8).

TABLE 4.8: A example of OCR results of the different submitted approaches in the case of complex background.

Method	SR image OCR results
Synchromedia Lab	 `Vofre_peou_vbus_parle`
ASRS	 `ya_'_WE`
SRCNN	 `m_4?_fix_E`
Groundtruth	 `Votre_peau_vous_parle`

4.5.4 Conclusion regarding the competition

The organisation of this competition was a challenge on several levels. First of all, a pertinent dataset had to be constructed. Although many text datasets have been proposed in the past for different purposes (detection, spotting, recognition, segmentation), the need of finely selected example was obvious, as they include images which initial form has poor resolution. They were therefore not suitable for SR task which requires HR images as references. In the same logic, a good OCR score had to be obtained on the HR images to clearly outline the interest of SR methods for LR text images. To match those criterion, the original HR examples (called "HD") were extracted from HD TV stream, and poor candidates were removed. A second challenge was to attract researcher among the SR community to this specific text-oriented task, and bring attention to the SR approaches among the document community. Although the number of participant was limited, figures of both community took part in the competition. The winners authors of famous neural-based publications on SR [DLHT14, DLHT16] also published a technical report [DZD+15].

The results of this competition indicate a general trend in favour of learning-based methods. This also supports the claim that neural networks are highly relevant for SR, even for non-natural images, and that they are now state of the art for such task. They also indicate that SR could be used in conjunction with other approach such as taking in consideration the foreground/background relationship to improve OCR performance on text with complex background.

However, we can see the limit of the proposed task as the winner reach scores that are very close to the original images. For a future competition, we would advise to further reduce the resolution of the input images to foster new methods that address SR using the context or relying on advanced models. Incorporating other tasks such as multiple scales or scripts different from latin could also provide a stimulating challenge and a good framework for experimenting novel approaches.

4.6 Analysis of the various learned priors

A last analysis that we propose to conclude this chapter is to observe the influence of the different learned priors. We propose to cross the model presented with various test data. This give a sense on how the model can generalize or handle new type of data.

4.6.1 Document text image

The *ULR-TextSISR-2013a* dataset is composed of black text over a white background. This bimodal nature has been illustrated in Figure 4.2. However, the text is generated with an anti-aliasing filter, that provide smooth edges for letters and is used in most digital displays to provide a nicer reading experience [GAF+87]. This means it contains also grayscale intensities value. For the LR and bicubic images, similar statistics are observed, but with more in-between values, as more black and white pixels have been merged together.

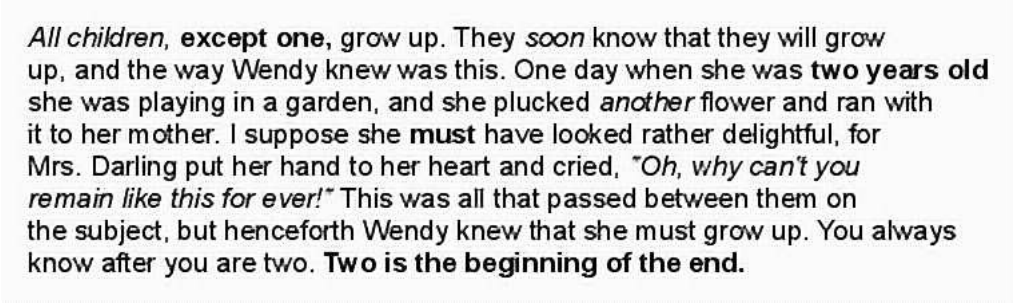
If a model learns on such data to generate directly the intensity values, the obtained images are expected to exhibit the statistical behaviour of the training data. To evaluate the impact of the training data, we train a model on the *ULR-TextSISR-2013a* dataset, and apply it to natural images and the *ICDAR2015-TextSR* dataset. An expectation might be to obtain an increased readability of the textual data of these datasets, under the condition that text characteristics (font family, size and style) lie close to the original training data. However, the sharpened images displayed on Figure 4.26c hold many artefacts as such data was not present during the training.

The differences are well outlined, notably by the difference in intensity. However, by sticking to the scheme we have initially adopted *i.e.* compensation of the higher frequency band by predicting the difference between the HR and the interpolated image, we obtain images which low frequencies are less impacted. However, similar strong responses on edges are observed and the "phantom noise" is still present.

4.6.2 Natural and TV Text Image

We can expect more subtle differences when comparing the priors learned on natural images and the TV-based text images from *ICDAR2015-TextSR* dataset. Complex backgrounds are present in the second, and they are from natural image content as they belong to the different TV shows which falls in the category of natural images. However, they do contain text from which we can expect to have a more specialised prior, even though the texts are less specific than the *ULR-TextSISR-2013a* dataset as they present a wider variety of fonts.

An illustration of the learned prior is to apply them to the *ULR-TextSISR-2013a* images. We observe in Figure 4.27 that different artefact appears. Results are much more smooth, which is in coherence with the smoothness priors used in many image restoration approaches. However, they tend to be oversmoothed, as sharp letters are expected in this context. We also notice overshooting (especially with our result in ICDAR2015),



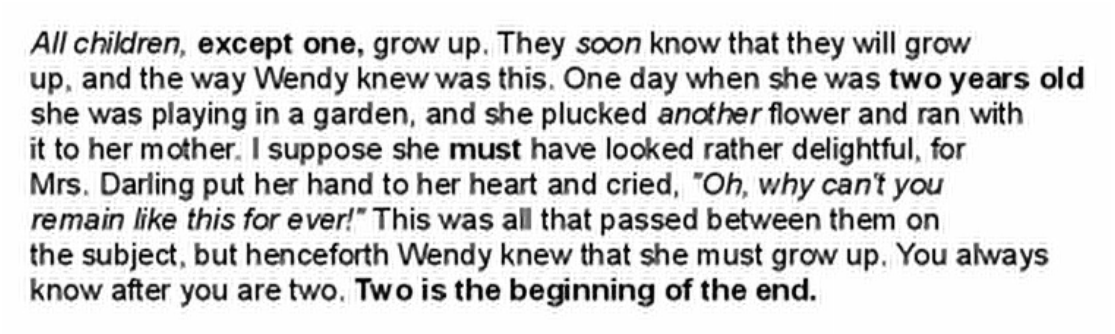
(A) Text (*ULR-TextSISR-2013a*)

(B) Natural Image

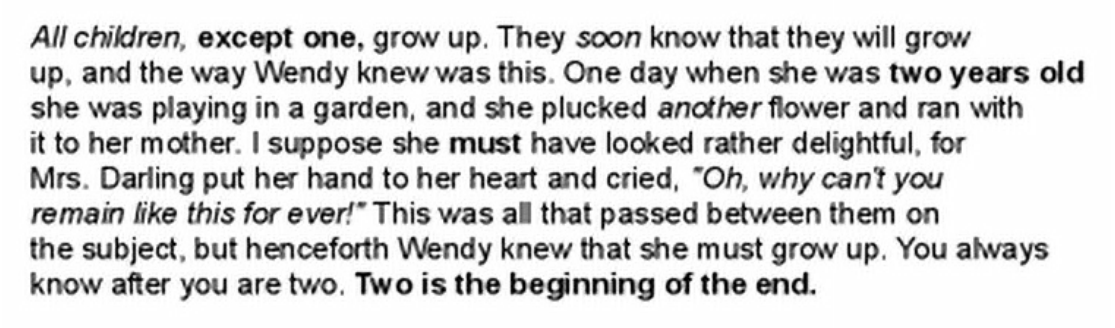
(C) Caption Image (*ICDAR2015-TextSR*)

FIGURE 4.26: Effect of the prior obtained with the *ulr-textsisr-2013a* dataset for graylevel prediction when applied to *ULR-TextSISR-2013a* dataset images, natural images and textual images datasets. While the *ULR-TextSISR-2013a* dataset test image in 4.26a is well shaped, the two others are over-sharpened and tend to have dissimilar dynamics compared with the expected high-resolution image.

which can be viewed as a smoother version of the observed noise reported previously (paragraph 4.4.3.2).



(A) Model trained on ICDAR2015 dataset
$PSNR = 20.21dB$, Accuracy= 91.28



(B) Model trained on natural images
$PSNR = 19.47dB$, Accuracy= 92.09

FIGURE 4.27: Results using the networks learned (A) on ICDAR2015 training data and (B) on natural images. The respective results in PSNR and Accuracy for the whole *ULR-TextSISR-2013a* testing dataset are displayed under each image.

4.7 Conclusion

We can draw several conclusions from this chapter on single image text super-resolution. First is that neural networks are suitable for Super-Resolution, as presented in early studies and widely explored in recent works. They provide an efficient framework to learn end-to-end mapping between LR and HR images. In particular, CNN offer better performance compared with MLP for the same complexity, due to improved non-linear feature extraction and mapping.

Images containing texts such as document images are severely degraded when sampled at low resolution. OCR engines performance on such images is decreased. However, they may be well reconstructed if such neural models are trained with suitable data. This data adaptation method allows to easily develop specialised SR systems. Moreover, without modifying the internal behaviour of an automatic recognition system, it is possible to improve its performance by feeding it with better shaped images. The proposed

ICDAR2015-TextSR dataset and competition also outline the interest of example-based SR approaches for other kind of text images such as text extracted from televisual streams. Very useful for multimedia indexation systems, they present a large diversity of fonts, colours, shapes and orientation. This public dataset can be used for evaluation of SR approaches, and provide an interesting double evaluation scheme based on reconstruction and OCR accuracy.

Finally, we outline two challenges that emerge from our study:

1. **Very low resolution text & Higher SR factors** – While the resolution of text images addressed in this study is already a lower bound of what can be usually addressed by recognition systems and human eye, we can expect to run into lower resolution in diverse situations. This already happens in some samples from the proposed ICDAR dataset, in which the height of letters may no exceed 3 pixels for the low-resolution versions. We have seen on the various proposed approaches that none of them yield satisfactory results on such images.

2. **Handle the variety of fonts & texts in the wild** – Another noticeable challenge lies in the nature of the texts present in everyday life. We addressed SR for text extracted from document images with the *ULR-TextSISR-2013a* dataset and TV content with the ICDAR dataset. Although the ICDAR dataset contains a good panel of fonts, they are still specific to typical TV news or advertisment. However, the variety of font that may be encountered in "real-life" contexts. Moreover, we only address latin script while there are many other scripts with various characteristics and different dependance on resolution. For instance, Arabic script contains meaningful informations based on small punctuation.

Chapter 5

Face Single Image Super-Resolution

Contents

5.1 Introduction

Facial images are useful for visual data understanding and automatic indexation. Indeed, the automated analysis can provide metadata about the people present in photos and videos: identity, age, gender, rate of appearance. While addressed for many decades, these research areas are still active today, as large-scale image datasets are available to research groups or companies through social networks or video sharing platforms.

However, resolution is again a bottleneck for such technology, as it is based on feature discrimination from a subject to another. Reduced resolution yields undistinguishable

components confusing human or automatic systems, which explains the interest for face super-resolution or face hallucination (see paragraph 3.4.2 of chapter 3). In a broader scope of application, increasing face resolution can be helpful for surveillance or improved customer and user experience.

For the multimedia indexing application we are interested in, face detection and recognition can lead to enriched navigation through contents by mean of a search including identity. Face clustering may be used to chapter a content or divide a scene into speaker turn. In this work we will focus on face recognition, although other technologies may benefit from improved resolution.

As shown in the literature review (paragraph 3.4.2 of chapter 3), two main SR approaches are adopted – sometimes jointly – and incorporate priors about the facial image in different ways:

1. *global approaches* in which the whole face is encoded in the LR space and recovered in the HR space
2. *local approaches* in which fixed regions of the face are processed independently, including adaptive approaches in which the face alignment is relaxed or included in the SR problem

Given the context that we want to address, faces are more likely to appear in non-aligned and uncommon modes. This leads us to present a method that falls in second category, with an adaptive scheme.

This chapter presents a two-step approach using neural networks. The first step is a generic local approach that improves the whole image resolution. The second step focuses on facial components (eyes, nose, mouth), with dedicated SR models. Each of these steps is described in section 5.2. Experimental results are presented in section 5.3, including a description of the data used in this chapter, architectures selection and experimentations that show the benefit of each step. Comments on the advantage and the limitations of the proposed approach are given in the conclusion, section 5.4.

This approach is the result of a joint work with Guillaume Berger, conducted during his 5 months internship at Orange Labs.

5.2 A two-step approach for face Super-Resolution

We propose to tackle the problem of face SR by adopting a two-step approach. The first step consists in improving the whole image resolution, with a similar data-driven

philosophy as the methods employed in chapter 4 for Text SR. The second consists in incorporating *domain-specific knowledge* about the data to further improve our SR system. This is summarized in Figure 5.1, and we detail each step in the following subsections.

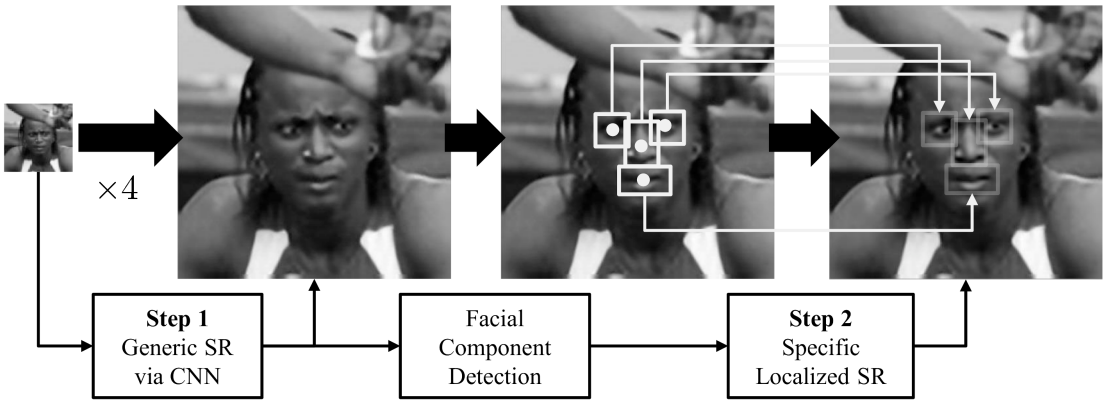


FIGURE 5.1: The proposed two-step neural approach for face SR. The first SR step is a generic one, that increases the resolution of the whole input image. The second step focuses on facial components and produces better shaped eyes, nose and mouth.

5.2.1 Step 1: Generic Super-Resolution

The first step of the proposed method consists in performing a first SR pass to improve the global image resolution. The objective is not to limit the resolution improvement to pose-specific face images (*e.g.* front aligned), but to begin with a generic SR pass, only specific to the nature of images it has been trained on.

Building on the proposed method for text, we have to adapt our data to the use-case and define a suitable ANN model. As facial image databases are available online but often for face recognition purposes, we propose to use one of those databases and adapt it to our study case. In the same way, pairs of low and high resolution patches are extracted to train the models. However, as facial images often have high resolution in their raw format, higher downsampling factors are investigated to address non-trivial and pertinent cases. This is explained in more details in subsection 5.3.2. Given the successful results obtained with the CNN presented in the previous chapter, similar models are investigated. Models description and experiments are reported in subsection 5.3.3.

However, face recognition involves *a priori* knowledge on the face structure and components (*e.g.* eyes, nose and mouth). As the present CNN is blind to these details, both due to the variety of samples in the training set and the limited size of the input retina, a second step is proposed to focus on these face components.

5.2.2 Step 2: Specific SR for facial components

The second step of the proposed approach aims to incorporate a domain-specific knowledge into the SR framework. Most low-level discriminative features between human faces are contained into facial components such as mouth, nose, eyes. Other factors might be taken into account such as hair, but are subject to higher variability within the same class (person). Other important high-level features such as the general shape of the face, proportions and components relative positions are also important but they do not suffer much from the downsampling process. We rely on the first pass to recover these features. However, given the importance of the first kind of features and the high probability they hold to be highly degraded in low-resolution images, a dedicated model is relevant.

Those features cannot be blindly sampled in face images, especially if they are taken in the wild, as they appear in specific locations. Thus, to incorporate this prior knowledge, we propose to exploit a facial landmarks detection algorithm (such as [BJKK13]). This allows to precisely extract facial components from the output image given by the first step, and train for each of them a specific model. The proposed architecture is described in paragraph 5.3.4.1.

5.3 Experimental results

5.3.1 Evaluation protocol

Similarly to text, an evaluation scheme must be provided to take into account reconstruction criterion and task-oriented evaluation. The standard PSNR measurement is used to evaluate the reconstruction performance of the proposed approach, and a face recognition score of an off-the-shelf recognition engine evaluates the impact of the SR algorithm on the recognition task.

We use a face verification system inspired by the simile classifier described in [KBBN09]. This system is a binary classifier that takes two faces as an input, and outputs a score characterising the similarity of the two input faces. Depending on the chosen threshold for this output score, the two faces are classified as belonging to the same or different person. Note that the performance of this off-the-shelf recognition engine is not competitive with current state of the art approaches. However, our goal is to evaluate the benefit of the proposed SR algorithm in terms of face recognition, which can be outlined with the chosen method (and more generally, with any other reasonably performing recognition engine).

The performance of the engine is characterized by ROC curves and the mean accuracy. ROC curves represent the true positive rate against the false positive rate, when varying the discriminative threshold between the two classes of a binary classifier. The mean accuracy is computed accordingly to a 10-fold validation: for each test fold, a threshold corresponding to 10% of false positive rate of the classifier score is found on the nine other folds and used to compute accuracy on the current one. The mean accuracy is obtained by averaging over the ten scores.

5.3.2 Data: Adapting LFW for Super-Resolution

Many databases have been released through the years in the domain of facial analysis. In the last years, database sizes have increased, with more and more available data and teams around the world to collect them. However, these datasets contain unconstrained facial poses. Three reasons can be found for this: the difficulty to have a large scale acquisition campaign with the same protocol, the availability of methods to align and normalise faces to reproduce a similar acquisition, and the need for data that is representative of real-world application.

Among those large-scale datasets, LFW [HRBLM07] (Labelled Faces in the Wild) has been very popular since 2007 for the study of facial analysis in unconstrained environment. It is relevant for this work as we focus on real-world application in the multimedia context.

FIGURE 5.2: Typical images from the LFW dataset. Faces are present in an unconstrained environment, spanning different poses, expressions, gender, ethnicity, and image quality.

Faces are present under different expressions, poses, expositions, illuminations and are sometimes partially occulted. We generate the LR images by blurring the original ones with a gaussian kernel of standard deviation $\sigma = 1.6$ and linearly downsampling them by a factor of 4. The choice of $\sigma = 1.6$ for the Gaussian kernel is a commonly used value [Sun08, YY13], and the use of linear downsamping is the simplest choice to reduce the

sampling rate while conserving the subpixels aligned (see subsection 2.2.2 of chapter 3). The downsampling factor is set to 4 as the images tend to have a large resolution in their original format. With factors of 2 and 3, no severe degradation of the performances on the dataset were observed. Starting from 240×240 cropped HR images, 60×60 LR images are obtained. The training images are selected so that they are not present in the testing subset, which gives $5,233$ images. $8,000$ images are used in the testing set, to form $6,000$ pairs for face recognition under ten different folds. From the generated LR and HR image pairs, we randomly extract 9×9 input patches with the corresponding 4×4 target ones and train the first step CNN. Then, for the second step, facial components are automatically extracted from the generic SR and HR images with an algorithm based on facial landmark detection [BJKK13].

5.3.3 First step

In the first step, we evaluate the impact of a generic SR based on CNN, simply adapted to the facial image case.

The 2D patches are extracted from face images taken in the wild, which are very close to natural images as they contain faces in various positions and a surrounding environment. The low and high resolution pairs of patches are blindly sampled without any knowledge on the location of the face. Therefore, the learned CNN is designed to be generic and learn to remove natural interpolation artefacts, via a global optimisation over all example pairs.

5.3.3.1 Architecture selection

To span a large scope of possible architectures, several experiments are conducted for the first step. Based on the previous experiment, CNN architectures with 3 and 4-layered networks are explored. As $\times 4$ SR is addressed, it seems legitimate to employ deeper architectures.

As with previous experiments, the CNN weights are learned with standard backpropagation and mean squared error loss function, taking as input 2D LR patches and targeting pixel-wise difference between HR and bicubic patches (see Figure 5.3). The latter corresponds to the loss of visual information and aim to compensate for artefacts such as blur or jagged edges. The variety of these artefacts, especially for high upscaling factors, makes the problem difficult and highly non-linear.

Table 5.1 gathers five tested configurations, where the $-a$ suffix indicates the connection scheme presented in paragraph 4.3.2.2, the -1 suffix stands for one to one connectivity,

and $-f$ for fully connected. The reported PSNR corresponds to a subset of 700 test images used to evaluate the performance of each configuration.

TABLE 5.1: Different deep architectures tested for the first, generic step. The $-a$ suffix indicates the connection scheme presented in section 4.3.2.2, the -1 suffix stands for one to one connectivity, and $-f$ for fully connected.

Configuration	1	2	3	4	5
Layer 1 (5×5 filters)	20	20	32	20	20
Layer 2 (3×3 filters)	230-a	230-a	560-a	230-a	230-a
Layer 3 (3×3 filters)	230-1	64-f	560-1	230-1	64-f
Layer 4 (1×1 filters)				64-f	64-f
Output Layer	16-f	16-f	16-f	16-f	16-f
Number of parameters	10,526	138,114	25,472	22,654	142,274
PSNR (dB)	35.162	35.324	35.169	***35.473***	**35.504**

Because configuration 4 has the second-best results but keeps the number of parameters low, it is selected for the rest of the experiments.

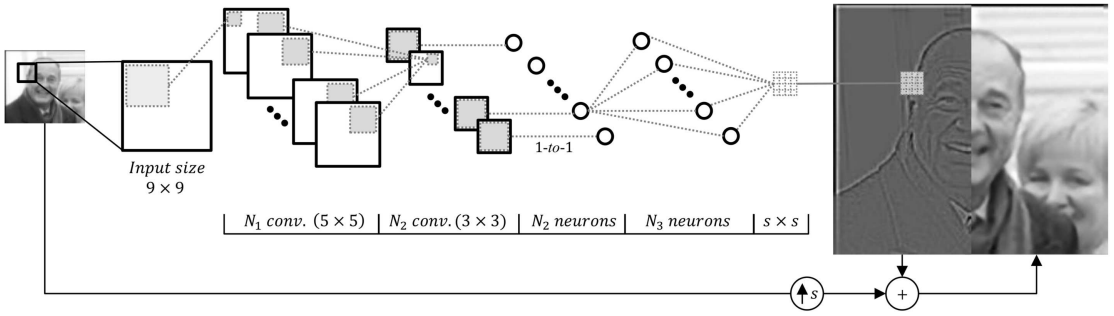


FIGURE 5.3: Selected CNN Architecture for the Generic SR step. Parameters are set to $N_1 = 20$, $N_2 = 230$, $N_3 = 64$ after testing different configurations, and $s = 4$ for the experiments. Note that the input image is still sampled on the LR grid, while the output map is sampled on the HR grid using $s \times s$ linear output neurons, yielding a s^2 times larger image.

5.3.3.2 Performance of the generic step

On the whole test set, an increase in the PSNR measurement is observed compared with the bicubic interpolation, which means that SR images lie closer to their original HR counterpart. The mean PSNR is $28.84dB$ for the bicubic interpolated images and $32.28dB$ for the SR images.

For recognition purpose, the selected architecture also leads to better results. The bicubic interpolation already allows to improve the recognition over the low-resolution

images score, in which faces are sometimes not even detected, and for which the features of the simile classifier seem badly extracted. The obtained SR images allow to further improve this score. By taking a false positive rate of 0.1, we obtain a mean accuracy of 81.61%.

In order to visualize the differences between the compared images, some samples from the testing set are depicted in Figures 5.4 and 5.5. Images have finer edges and reduced LR artefacts. Some textures are also well recovered. However, details of facial components are sometimes lost completely (*e.g.* teeth in Figure 5.4).

The second step aims to compensate for this loss of details by mapping the obtained SR facial features to an improved SR version using dedicated models.

5.3.4 Second step

5.3.4.1 Autoencoder architectures for component-specific models

For this second step, we aim to improve the blindly reconstructed facial features (eyes, nose, mouth). To do so, four different networks are used – one for each facial feature – with the same architecture, depicted in Figure 5.6.

For each facial component, a convolutional encoder projects the input patch into a N_D-dimensional hidden subspace, and the output patch is reconstructed from the obtained code through a fully connected one-layer decoder with linear activation. The output patch can then be written as a weighted linear combination of N_D atoms c_n:

$$o = \sum_{n=1}^{N_D} w_n \cdot c_n \tag{5.1}$$

where o is the reconstructed output facial component, w_n are the components of the code in the hidden subspace, and c_n are the weights of the decoder layer associated to the n^{th} code component. Each atom is directly associated to one direction of the hidden subspace. In order to learn a meaningful representation, a sigmoid activation is added on the encoder output to make the code positive, and a non-negativity constraint on the decoder weights. As presented in Figure 5.6, this constraint makes atoms c_k become part-based and less noisy, similar to non-negative matrix factorization [LS99].

5.3.4.2 Evolution of the performance compared with the first step

Decreased performance for compliant reconstruction The PSNR suffers from the second step as mean PSNR of $31.64dB$ is obtained, against $32.28dB$ for the first

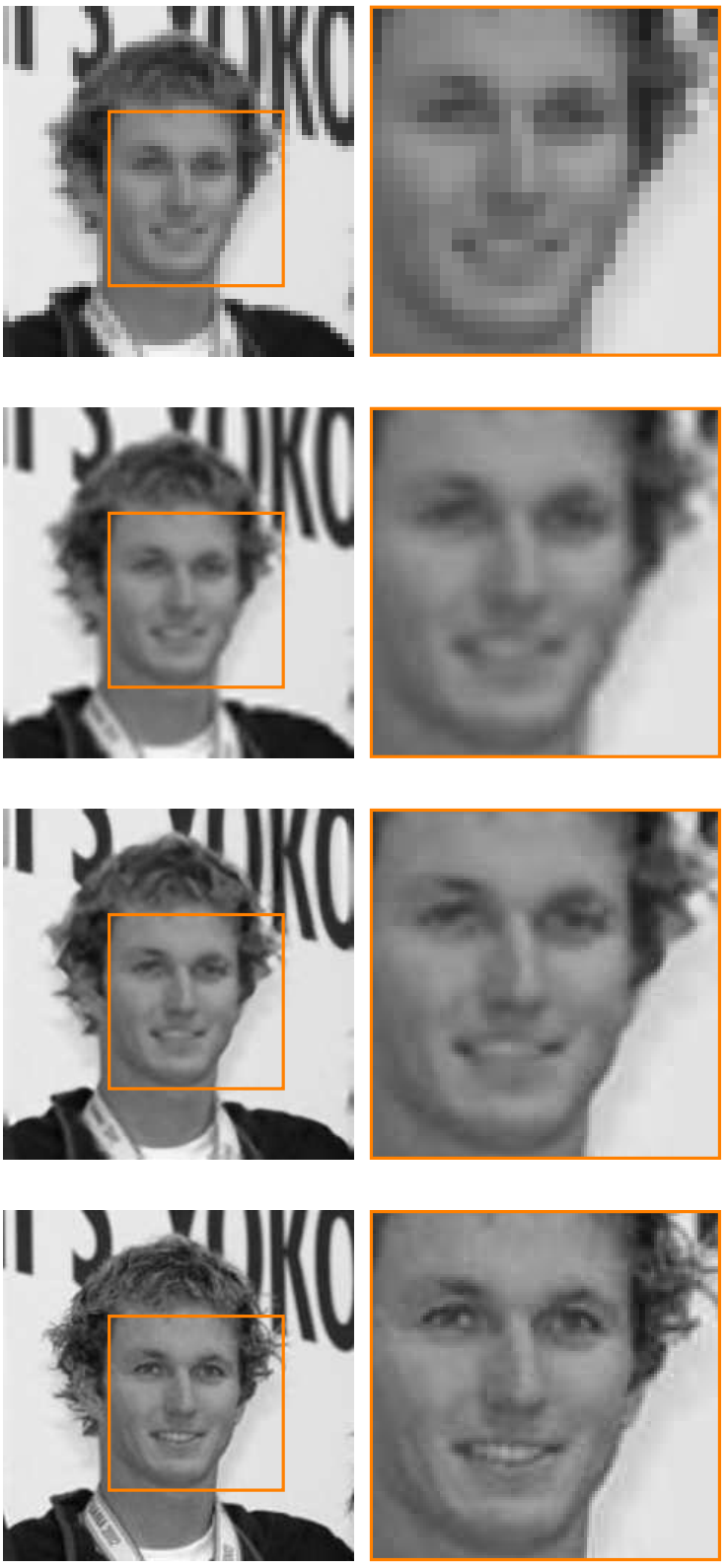


FIGURE 5.4: "Aaron_Peirsol_0002" picture, from top to bottom: LR image, Bicubic interpolation, Results of the first step, and original HR image. Edges are globally well reconstructed, without blur or jaggy edges. Textures such as hair is also finer, but they lack of realism compared with the original image. The facial features also exhibit severe damages even if sharper.

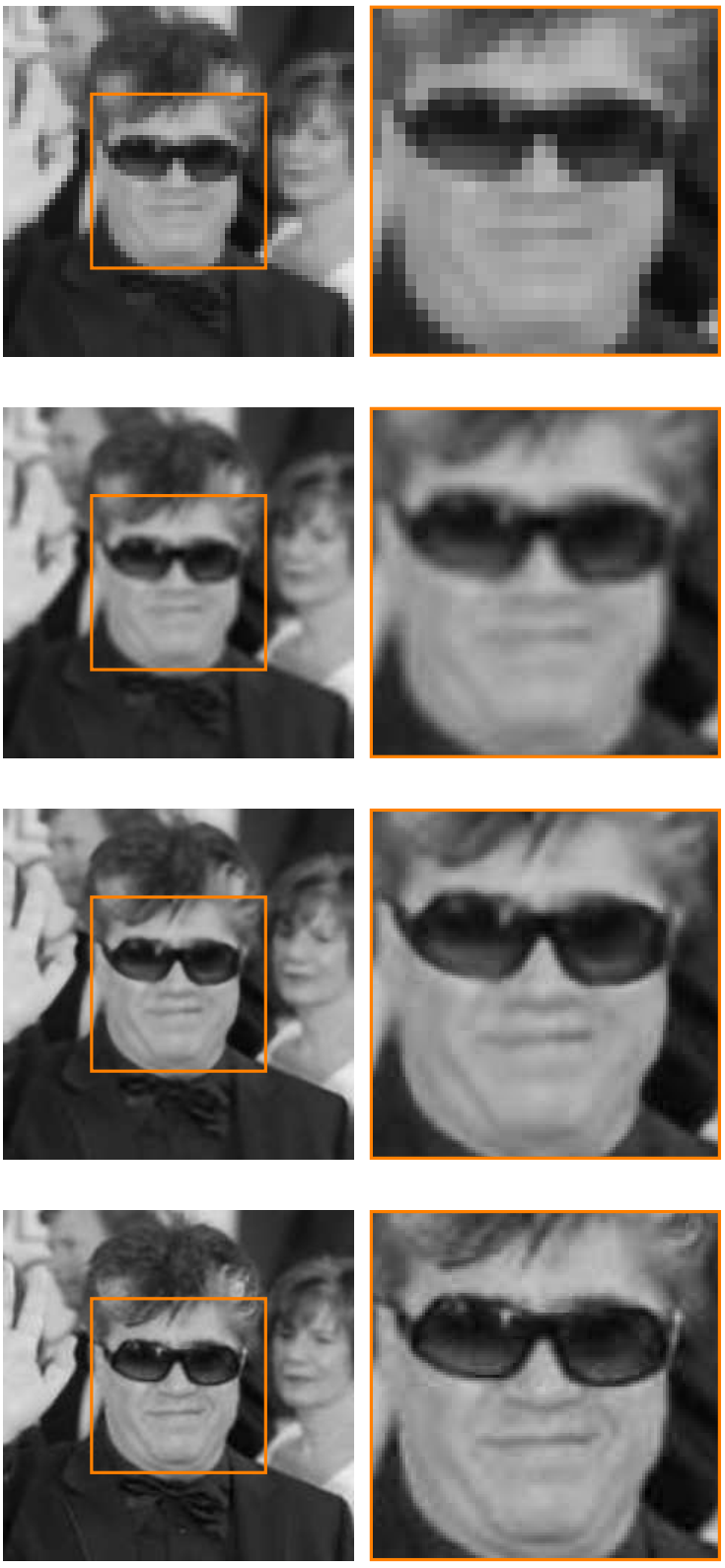


FIGURE 5.5: "Pedro_Almodovar_0003" picture, from top to bottom: LR image, Bicubic interpolation, Results of the first step, and original HR image. Again, edges are well reconstructed (particularly sharp on the glasses border), demonstrating the ability to address bigger upscaling factor with the proposed method. However, a fine reconstruction of the facial features is lacking.

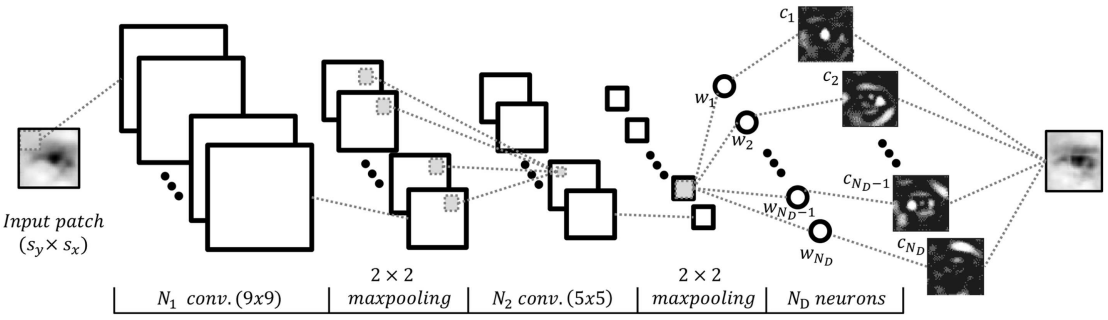


FIGURE 5.6: Localized models for facial components. Left and right eyes, noses and mouths are extracted and processed by distinct networks. For eyes and nose patches: $s_x \times s_y = 36 \times 36$. For mouth patches: $s_x \times s_y = 48 \times 24$. Reported results were obtained with $N_1 = 8$, $N_2 = 64$, $N_D = 128$.

step. This diminution can be explained by the fact that positivity constraints added on the code and decoder weights make the reconstruction goal harder to fulfil: second step outputs have to be produced by adding a limited number of positive atoms. As a consequence, facial components given by localized models tend to differ from HR targets and make the PSNR drop slightly. However, even if they are different from the target, facial components given by localized models contain less visual artefacts which were still present after the first step. This is mainly explained by the fact that the second step outputs are reconstructed by combining clean part-based atoms.

Better recognition While a PSNR decrease is observed, the specific models allow to obtain a better recognition score, with an additional 1.24 points (82.85%) over the test set compared with the first step (81.61%). The ROC curves reported in Figure 5.7 illustrates the recognition performance for a varying threshold in the recognition engine for the LR, bicubic, SR (step 1), SR (step 1+2) and HR images. Table 5.2 and Figure 5.8 summarise the obtained score for the different category of images.

TABLE 5.2: Results of the proposed 2-step approach on the LFW corpus. The first generic step allows to improve PSNR and accuracy by producing a ×4 SR image. The second specific step on facial components slightly reduces the PSNR as the produced image is not exactly compliant with the original HR image, but further improves the accuracy.

	PSNR (dB)	accuracy (%)
LR	-	74.70
Bicubic	28.84	78.91
SR - step 1	**32.28**	81.61
SR - step 2	31.64	**82.85**
HR	-	86.55

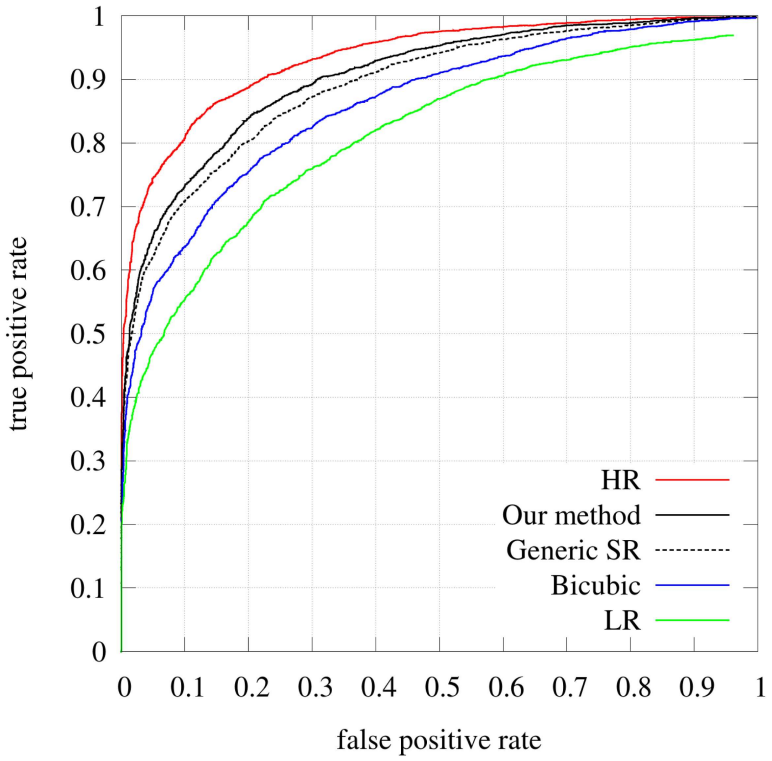


FIGURE 5.7: ROC curves illustrating the performance of recognition of the recognition engine given the different image sets.

5.3.4.3 Other observations

Side effects can be noticed on the images produced by the second step. First of all, it tends to hallucinate normalised facial components, which has good aspects outlined in the samples of Figure 5.9 and the substantial increase in recognition score, but can lack of compliance with the original images. For example in Figure 5.10, the make-up around the eyes is sharp but too present after the first step, and removed by the second step with more realistic but less accurate eyes contours. The second step may also hallucinate components where they might not be present in the original image. In Figure 5.10 "Pedro_Almodovar_0003", the eyes are revealed ("hallucinated") and merged with the sun glasses, which is not compliant with the original image.

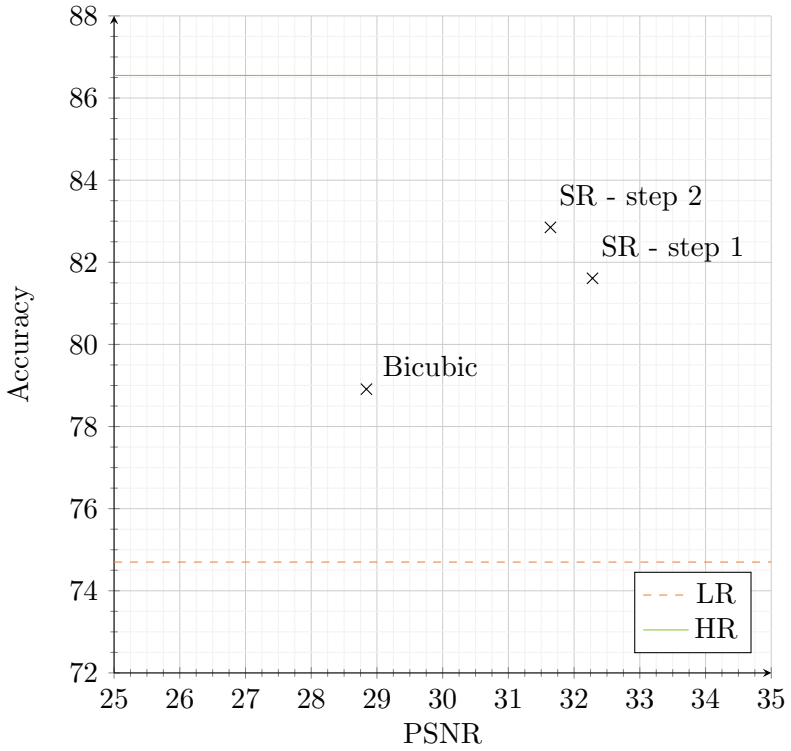


FIGURE 5.8: Summary of the obtained results. Both steps allow to improve reconstruction and recognition over a simple bicubic interpolation. The first step is more efficient for reconstruction (PSNR) as the facial component are not hallucinated but reconstructed with the same *a priori* that the rest of the image. The second step has a slightly lower reconstruction PSNR score, but its specific models that focus on facial components allow to produce more realistic characteristics and improve recognition performance of the recognition engine. PSNR is not relevant for the LR (not the same dimensions) and the HR (infinite) images performance.

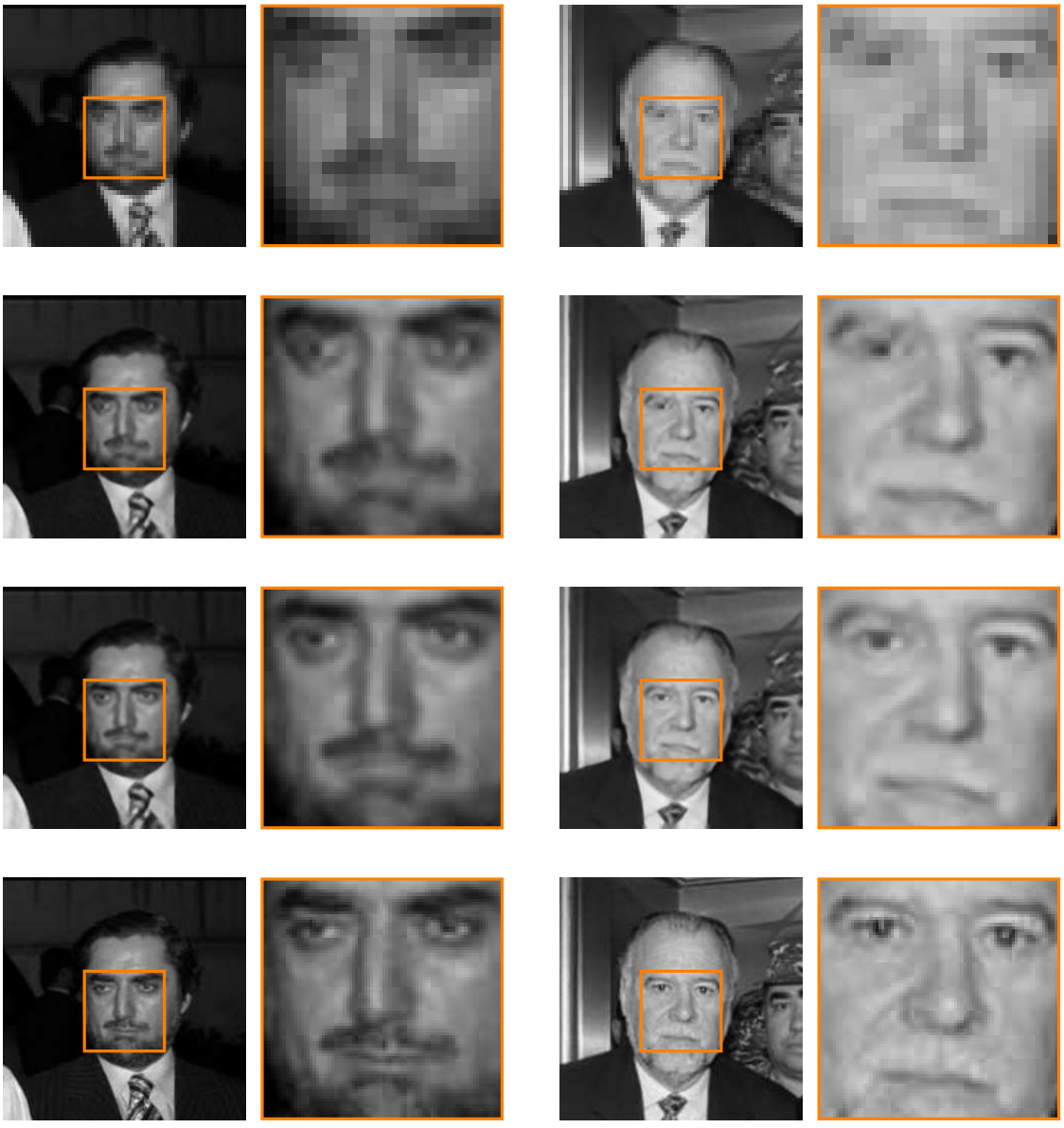
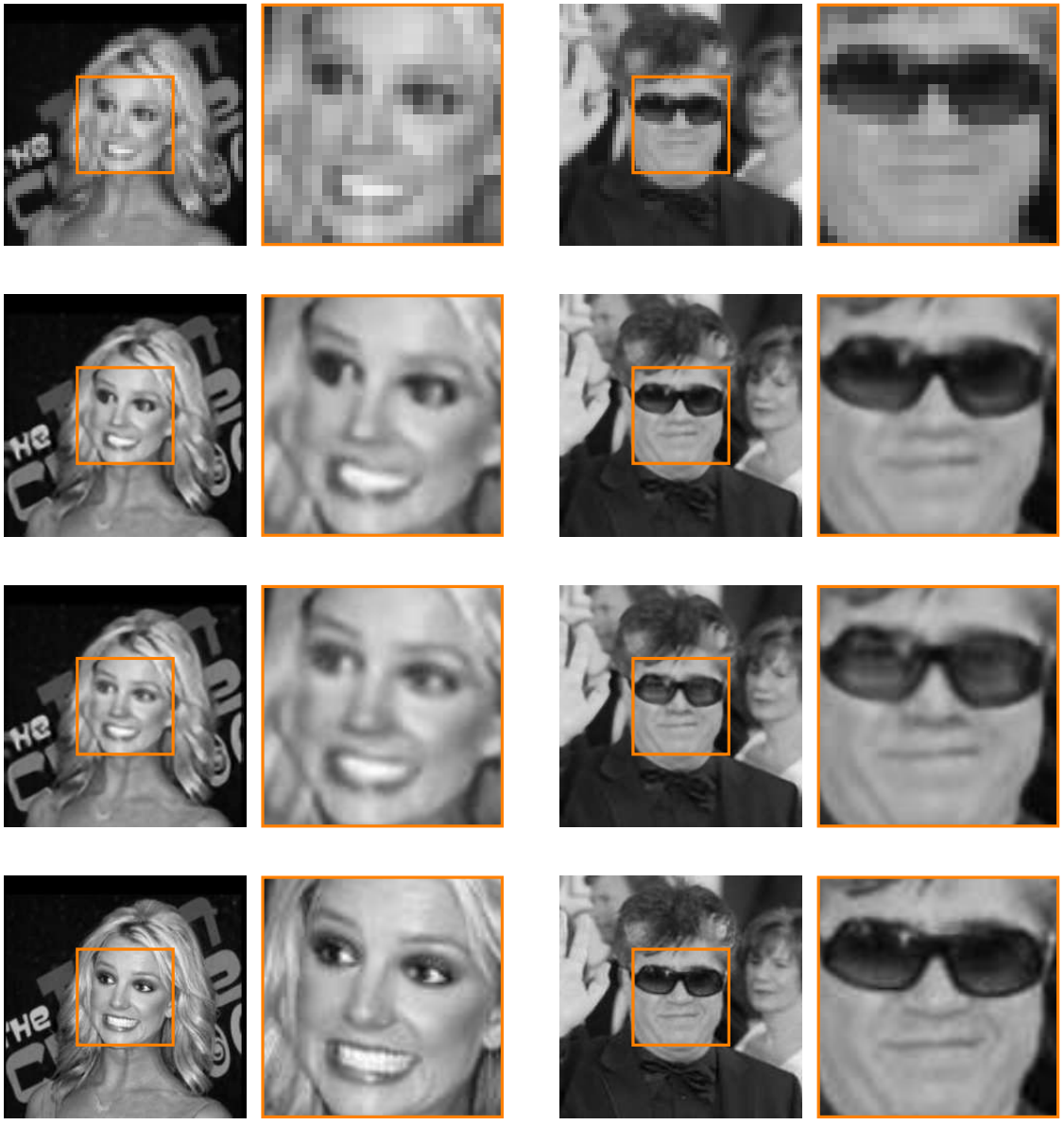


FIGURE 5.9: Results obtained with the proposed approach. from top to bottom: LR image, Bicubic interpolation, Results of the first step, Results of the second step and original HR image.



(A) "Britney_Spears_0012" (B) "Pedro_Almodovar_0003"

FIGURE 5.10: Side effects: hallucinated facial components may lack of compliance with the ideal HR image, even if improving the overall performance. From top to bottom: LR, SR (step 1), SR (step2), and original HR image.

(A) HR (B) Bicubic (×4)

(C) HR (D) Bicubic (×4)

FIGURE 5.11: Some images of the LFW dataset (here, "Abdoulaye_Wade_0003" and "Ahmed_Ghazi_0001") have low-resolution and contain compression artefacts. There is not much difference between the original HR image and the ×4 bicubic interpolation.

5.4 Conclusion

In this chapter, the method proposed for Text SR is extended to a new type of images and with a similar objective: producing better shaped and more recognizable facial images. However, due to the complexity of face geometry and the diversity present in the large scale dataset, the proposed approach decomposes the SR problem in two steps. The first one aims to produce SR images with recovered details, using a patch-based approach similarly to the previous chapter. In the second step, the focus is set on the facial features to train specific models for each of them. Using a facial landmarks detector, the image is further transformed to produce better shaped eyes, mouth and nose for each face. Even though the final image is not as close to the original image as after the first step, it allows to produce realistic facial components and improve the performance of the off-the-shelf recognition engine.

The proposed model could benefit from further improvement. First of all, a finer selection of the training data could be beneficial. Indeed, some of the high-resolution images from the LFW database already contain artefacts such as JPEG compression blocks, or obtained themselves from interpolation of a lower resolution images. For those images, there is no benefit from incorporating them into our learning set. This remark holds for both steps of the method as they are cascaded with the same HR training data. Also, our model requires several steps that are performed iteratively. Although they are fully automated, a unified architecture that incorporate the facial detection at its first stage might avoid having five separate neural models (1 for the generic step and 4 for the components).

Chapter 6

Blind and Robust Super-Resolution

Contents

6.1 Introduction

Recall the assumption so far that low-resolution images are generated from high-resolution ones via a fixed observation model (see subsection 2.2.3 of chapter 3). However, this may not hold in real-world situations, where the true observation model can vary and where other factors may be involved, relative to the environment, objects or devices used for image acquisition. This dependency is particularly hard to avoid for example-based methods which rely on coherent examples in order to produce accurate images. If only a fixed observation model is used to create the training images, such methods will be very efficient on this particular kind of image but with lack of generalisation on others, which occurs in real-world situations.

In this chapter, we investigate on making example-based SR approaches robust to the variability of observation model, in a *blind* way (*i.e.* without knowledge of the observation model). In particular, we want to provide robust solutions to the variability of the blurring kernel used in the observation model.

In order to benefit from the potential of learning-based approaches, three main strategies are considered in the literature for blind SR (see below, subsection 6.2.2): i) projecting the LR images into the known LR space – or alternatively predicting the right model to use from a collection – to fit the distribution of image seen during training ii) online retuning a pre-trained model iii) designing a learning machine able to implicitly model the projections into an end-to-end framework. The latter is a more relaxed problem for constrained image category such as faces or text. Throughout this chapter, we investigate the third strategy by using deeper CNN in order to try to absorb the different blurring kernels. With large number of parameters and highly non-linear projections, such models may have the potential to absorb the variability of the data and produce accurate SR images. Recent work on blind deblurring [Cha16] show that neural networks can address such problems. Moreover, experiments from [RSRB15] indicate that using a blind CNN already brings a gain over a simple bicubic interpolation.

This chapter begins with a discussion on the relevance of image observation model for real-world applications in section 6.2. A short review of example-based methods that aim to overcome those limitations is also given. In section 6.3, to better state the problem addressed by this contribution, measurement are made to better define the limits between a deblurring and a SR problem. Preliminary experiments on three different strategies to incorporate the variability of the observation model are presented in section 6.4. In section 6.5, a deep CNN approach is proposed to tackle the blind SR problem, with many blurring kernels added into the training set. The experimental

results are reported in section 6.6 and show that the proposed blind model can have similar performance as the non-blind models of the literature.

6.2 Discussion on the robustness of example-based approaches

6.2.1 Confronting the observation model with real-world conditions

This section discusses the relevance of the conducted work regarding real-world conditions, where a single image observation model (*i.e.* how do we obtain a LR image from a HR one) cannot represent the variety of cases (devices used, atmospheric conditions, motion). Recall for SISR, the following model was used:

$$y = DBx + e \tag{6.1}$$

where D is a down-sampling operator, B is a blurring operator and e is a noise term.

Noise In the scope of this work, the considered images are most of the time subject to low noise conditions. Therefore, we do not address nor expect high levels of noise. Indeed, in the first contribution in chapter 4, two types of text images have been considered. The first one is typically obtained from digital low-resolution documents (*e.g.* downsampled version of a digital document) or in good illumination conditions (such as documents obtained using a scanner). For this type of images, the assumption of low amount of noise is reasonable, unless images are highly compressed (*e.g.* JPEG artefacts) or inherent to the document (damaged or stained document). This was not the case for the used document dataset, and the HR images in the proposed ICDAR2015-TextSR dataset are obtained from downsampled HD ones which remove the potential noise. The second type of text images are extracted from TV streams for which bicubic downsampling was a sufficient approximation. For face images in chapter 5, the main problem is the dimension of faces and the loss of detailed facial components. However, even with those assumptions, these previous experimentations are not completely noise-free, as the images undergo quantization processes during saving, in which a transformed image (e.g. a newly synthesized LR image) with float pixel intensity values is mapped to a discrete valued image with 8-bit (256) values per channel. The quantization function

used in this work is:

$$i_{quantized} = \begin{cases} floor(i + 0.5) & \text{if } i \in [0, 255] \\ 0 & \text{if } i < 0 \\ 255 & \text{if } i > 255 \end{cases} \tag{6.2}$$

Even if noise is not addressed here as we consider it negligible compared with the problem of low resolution, it is still relevant in certain contexts. Many works of the literature handle noise, often as independently identically distributed Gaussian noise. Other common noise is compression and encoding artefacts, as most of images in the wild (web, phones) are compressed. The approaches focussing on this kind of noise can be utilized to first process the LR image before using the methods proposed so far.

Blur The approaches proposed in the previous chapters make use of two kinds of image observation models. In chapter 4, antialised bicubic downsampling was used to downsample the images by a factor of two. In chapter 5, a Gaussian blurring kernel of standard deviation of 1.6 (see subsection 5.3.2) and linear downsampling are used to downsample the facial image by a factor of four. These observation models are usual regarding the literature. However, they indicate SR is not decorrelated from deblurring, as a blurring kernel is part of the observation model. As a preliminary study, we will review in section 6.3 the cases for which it is relevant to address the upsampling problem as a SR problem rather than a deblurring one on a denser sampled grid. When looking at the imaging model in deblurring and SR, the main difference is the presence of the decimation operator. This means that depending on this operator, it may end up in pure deblurring conditions if no information is lost during the downsampling. In the frequency domain, sampling results in repeating the spectrum every ω_s (see Figure 2.2 in chapter 2.2.2). If the signal being sampled is not compliant with Shannon conditions [Sha49], aliasing will occur. Aliasing is a fundamental property in both MISR and SISR. In MISR, it guaranties that a non-redundant information have been split in different observations. For SISR, it ensures that the problem is actually different from a deblurring one. Many works (including ours in chapter 5) use Gaussian blurring kernel. Strating from this model, the limits between the SR and deblurring problematics ar evaluated in section 6.3.

6.2.2 Short review of blind approaches in example-based SR

Blind SR refers to methods that have no a priori on the used image formation model, especially the blurring kernel. Practically, a majority of the blind methods estimate the most probable blurring kernel and/or use statistics about the desired SR images. In this context, iterative and MAP approaches are often preferred, while direct example-based methods are not the most popular approaches. Recent works [YMY14, EGA+13] show the importance of knowing a precise blurring kernel in the learned prior by example-based approaches. When evaluating such methods without retraining the models with the right kernel, results exhibit over-smoothed or over-sharpened images: evaluating on non-blind models gives much better results. In [EGA+13], the authors study the influence of a Gaussian blur kernel with variable standard deviation applied before an antialiased bicubic downsampling. They show that a simple prior on the SR image may already produce good results as long as the exact image formation process is known. In [YMY14] an extended study compares recent learning-based approaches. Both studies show that retraining example-based approaches with the right kernel gives better results. In [MI13], a joint estimation of the blurring kernel and the SR image is performed. Later in [SE15], authors address both SR and deblurring, using the output of a learning-based SR method as a constraint on the final image. A global MAP optimisation scheme is then applied along with other deblurring constraints on SR and LR images. This unified approach is very interesting as blur is one noticeable artifact in interpolated images, along with jagged edges and other kind of noise. In [ZWTZ16], the authors combine blur kernel estimation and per-image dictionary learning, which is more precise but also slower. In [RSRB15], the authors proposed a richer collection of blurring kernels using oriented bivariate gaussian ones. Although the main purpose is to propose adaptive scheme in a non-blind fashion, they conduct the so-called blind experiments with several example-based algorithms.

Another problem that can relate to blind SR is the unknown scale problem, in which the scale factor is not known in advance. Cascading approaches have been proposed, like in [WLY+15] where the same model can be used for $\times 2$ and $\times 4$ SR. In [WYW+15] and [KLL15a], different scales are used for data augmentation that enhance the performance of a neural network by learning more robust convolution kernels in [WYW+15] or being blind across the scales for [KLL15a]. In [ZFC+15], the learned hallucinating CNN is also blind and robust to several blur kernel and resolution. However, it is likely constrained to aligned faces, which allows a strong prior on the type of output data.

6.3 Blurry or Low-Resolution ? Preliminary reflection on the observation model

Most of the observation models used in the literature include a low-pass filter (often Gaussian, Box, Bicubic) followed by a decimation. In particular, Gaussian filters are controlled by their standard deviation (or variance), which can be set to different values. However, if the low-pass filtering is such that no level of aliasing is contained in the low-resolution, the problem can be modelled as a deblurring one instead of a SR one. Indeed, the increase in resolution could be achieved optimally with a perfect reconstruction from the samples. In other words, the lack of *definition* (no high frequency content above the Nyquist limit) is already present in the image before it is decimated, and the inverse problem is well-posed.

To illustrate this, the impact of blur and downsampling on a set of 91 images is measured. A Gaussian blurring kernel with a standard deviation σ between 0.3 and 8.0 (by 0.1 step) is applied, followed by a bicubic downsampling operator. The impact of the different observation models thus formed is evaluated using the PSNR between the interpolated LR images and the original HR one. Figure 6.1 shows the mean PSNR for different downsampling factors depending on the σ of the blurring kernel.

We can see that while the PSNR measurements of each downsampling factor are clearly distinguished for small standard deviation σ, they get closer as the σ increases, *i.e.* the blurring operator removes a lot of the high frequencies. At a certain level, there is not much difference between the PSNR values, independently from the downsampling factor. Based on this measure, a standard deviation as small as 1.2 for a downsampling factor of 2 yields a problem as difficult as a decimation factor of 3 without a Gaussian smoothing. Considering that a difference of $0.1dB$ between blurry and LR images is a threshold under which they are subject to an equivalent level of degradation, each scale can be attributed a "maximum" standard deviation (see Table 6.1).

TABLE 6.1: Maximum standard deviation σ for each scale above which mean PSNR difference between LR and blurry images is less than $0.1db$.

scale	2	3	4	5	6	7	8
SR limit for σ	2.4	3.4	4.4	5.4	6.4	7.4	8+

In conclusion, this short study allows to qualify the boundaries of the SR problem versus a deblurring one. In the next sections, we explore ways to make neural-based approaches more robust to various observation models that may occur in real-world situation. We ensure that the spanned kernels are compliant with the previous study.

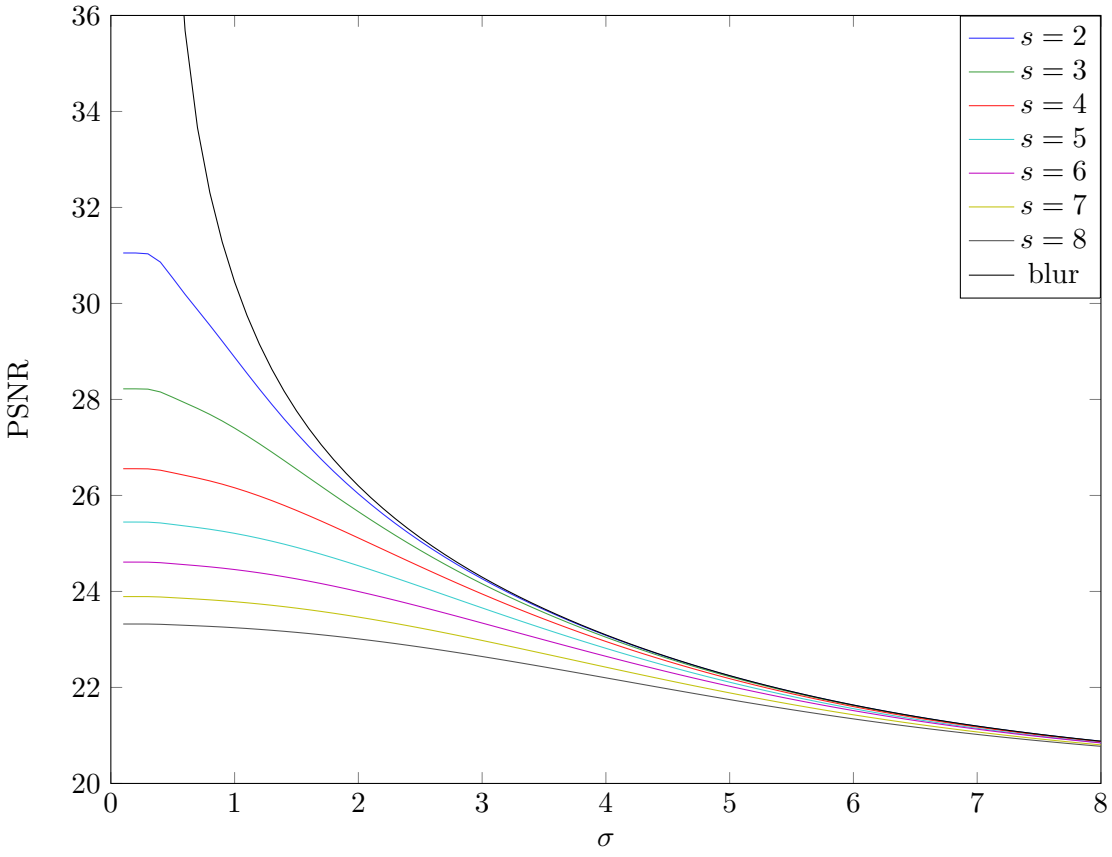


FIGURE 6.1: Evolution of the mean PSNR of the image set with the Gaussian kernel standard deviation for 7 different downsampling factors (2 to 8). The black curve is the PSNR of the blurry images without downsampling.

6.4 Preliminary 2-kernels experiments

Based on the previous insights, we propose to conduct preliminary experiments to explore the possibility for the studied learning systems to gain robustness against the different observation models that might be encounter in real world applications. To do so, two different Gaussian blurring kernels are used to form distinct observation models for a factor of 2, with standard deviation of $\sigma_1 = 1.0$ and $\sigma_2 = 2.0$, that match the range proposed in the previous section (see Table 6.1). Figure 6.2 illustrates this process. It aims to study the behaviour of the training process, and obtain insights on the behaviour of a neural network addressing different blurring kernels in the same model. Later in sections 6.5 and 6.6, we shall span a larger variety of kernels to address more realistic cases. Three different strategies are evaluated. The first one called "exclusive" is to use a model trained on single kernel generated data, and use it on different data. The second one called "fine-tuning" consists in training on a first single kernel training set and fine tuning on a second set with a different observation model. Finally, the third strategy called "inclusive" is to mix the training datasets and train a single model indifferently with the joint data.

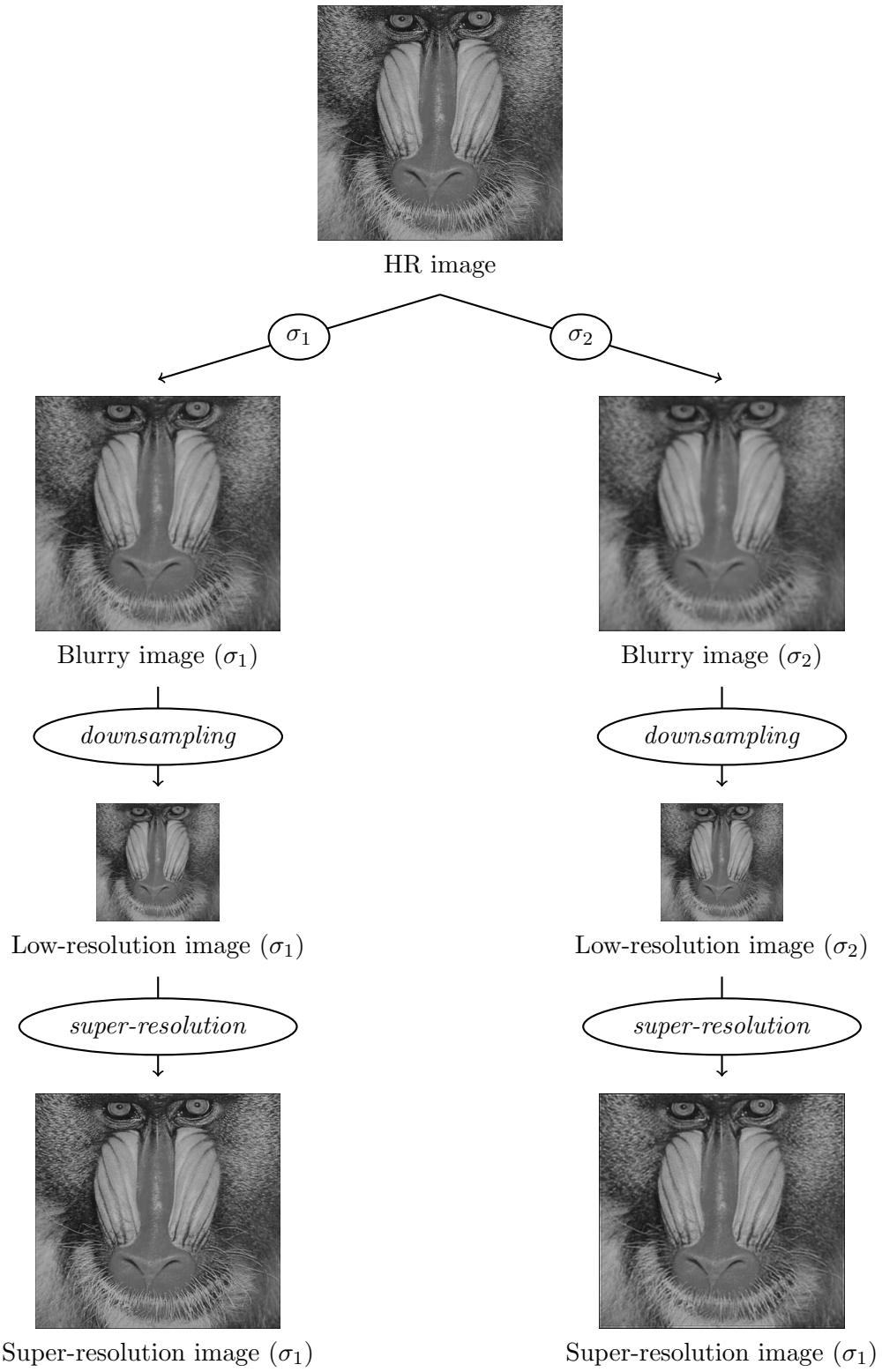


FIGURE 6.2: Two different blurring kernels are used to generate two types of LR images. A robust SR algorithm would be able to render the same SR image in both cases.

For evaluation, the PSNR criterion is used on a test set proposed in [YWMH08] composed of two subsets of natural images (*Set5* and *Set14*). Visual results are also examined. The reference is bicubic interpolation.

6.4.1 Exclusive training sets

In this first experiment, one model is trained for each dataset and evaluated on two distinct test sets. It allows to evaluate how the models behave on unseen data, but also illustrates the importance of the observation model for a learning-based system.

The results are reported in Table 6.2 and examples from the test set can be seen in Figure 6.3. As expected, each model perform well on the images generated with the same kernel as those in the training set. However, over-smooth or overshoot artefact are observed in the images generated with the unseen kernel.

TABLE 6.2: Results with exclusive training sets.

	$PSNR_{\sigma=1.0}(dB)$	$PSNR_{\sigma=2.0}(dB)$
Bicubic	28.97	26.13
Trained with $\sigma = 1.0$	32.83	27.02
Trained with $\sigma = 2.0$	20.06	30.04

6.4.2 Fine-tuning

The second experiment consists in two steps for each model. First, a model is trained on a set produced with a single imaging model. Then, the same model is fine tuned using new data to see if the neural network can be taught a new kernel for SR without forgetting the old one.

Those two steps are performed on the same data as before. The obtained results are reported in Table 6.3.

TABLE 6.3: Results with the fine-tuning strategy

	$PSNR_{\sigma=1.0}(dB)$	$PSNR_{\sigma=2.0}(dB)$
Bicubic	28.97	26.13
Trained with $\sigma = 1.0$ fine-tuned with $\sigma = 2.0$	18.98	30.38
Trained with $\sigma = 2.0$ fine-tuned with $\sigma = 1.0$	32.83	27.03

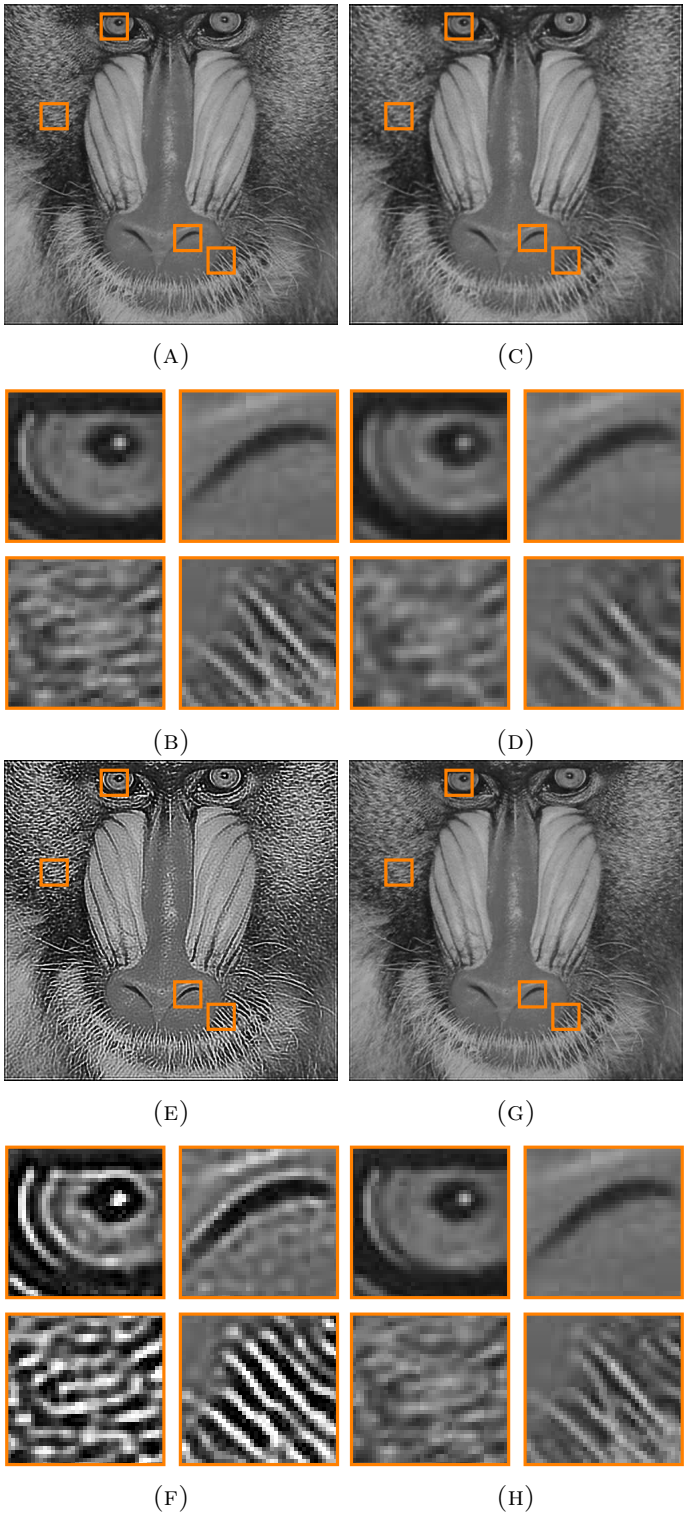


FIGURE 6.3: Test images and close-ups obtained with the first strategy (exclusive training sets), applying models indifferently from their learning data. A–D: Results of model trained with $\sigma = 1.0$, for LR images generated with $\sigma = 1.0$ (A, B) or $\sigma = 2.0$ (C, D). E–F: Results of model trained with $\sigma = 2.0$, for LR images generated with $\sigma = 1.0$ (E, F) or $\sigma = 2.0$ (G, H).

We can see that the strategy does not promote a consensus between both training sets, and the performance is approximately the same. Taking a look at the produced images in Figure 6.4, it is likely that the networks converge to a stable state which is uncorrelated with the one reached after the first step.

6.4.3 Inclusive training set

The last experiment consists in a fusion of both training sets. While the previous experiment (strategy 2, fine-tuning) aimed to capitalize on a memory of the learned weights – which turned out to be hard to hold – here the weights are continuously updated to minimise the prediction errors over the whole dataset, composed of LR images obtained with $\sigma = 1.0$ *and* $\sigma = 2.0$. This is a straightforward approach in which the network has no knowledge of which kernels are present in the training set. It must be able to absorb those differences by learning a richer set of filters.

In Table 6.4, we observe improved scores over bicubic interpolation. However, the improvement is less important than the one observed when training or fine-tuning with the $\sigma = 1.0$ dataset, yielding blurry images for the case of $\sigma = 2.0$ (see Figure 6.5).

TABLE 6.4: Results with inclusive training dataset.

	$PSNR_{\sigma=1.0}(dB)$	$PSNR_{\sigma=2.0}(dB)$
Bicubic	28.97	26.13
Trained with both $\sigma = 1.0$ and $\sigma = 2.0$	31.77	26.56

6.4.4 Conclusion of the preliminary experiments

This first study gives us several insights:

1. As stated in previous studies, learning-based approaches such as the one proposed are limited to the nature of data they are learning from since they do not explicitly represent the observation model. In the case of SR, if a single observation model is used to synthesise a dataset, performance is limited to this kernel.

2. For our two-kernels case, the fine-tuning approach seems impractical: convergence is hard to reach, and the balance between the base model and the fine-tuned one is delicate to obtain.

3. The straightforward inclusive approach is more promising, but seems to sometimes neglect parts of the data for the used model.

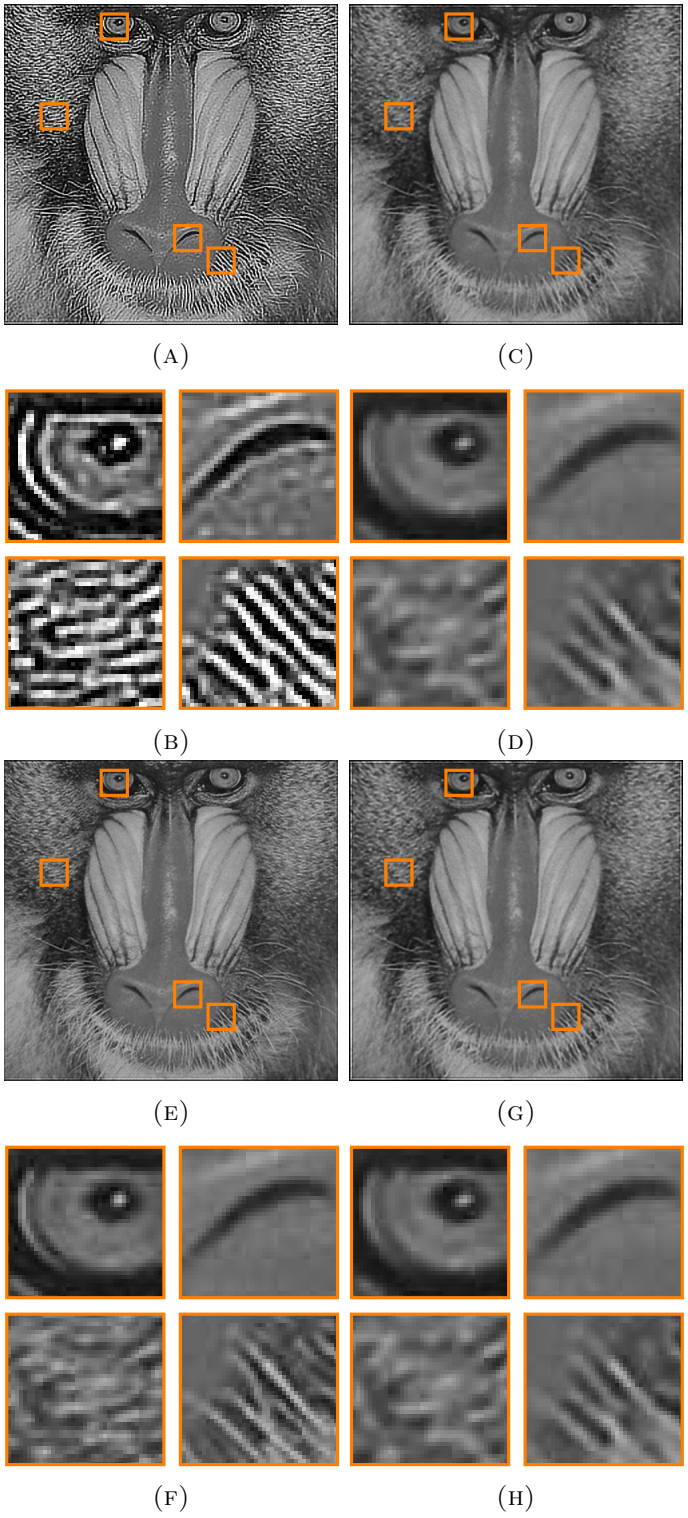


FIGURE 6.4: Test images obtained with the fine-tuning strategy. The networks tend to forget the state reached after the first stage and converge accordingly to the fine-tuning data. A–D: Results of model trained with $\sigma = 1.0$ and fine-tuned with $\sigma = 2.0$, for LR images generated with $\sigma = 1.0$ (A, B) or $\sigma = 2.0$ (C, D). E–F: Results of model trained with $\sigma = 2.0$ and fine-tuned with $\sigma = 1.0$, for LR images generated with $\sigma = 1.0$ (E, F) or $\sigma = 2.0$ (G, H).

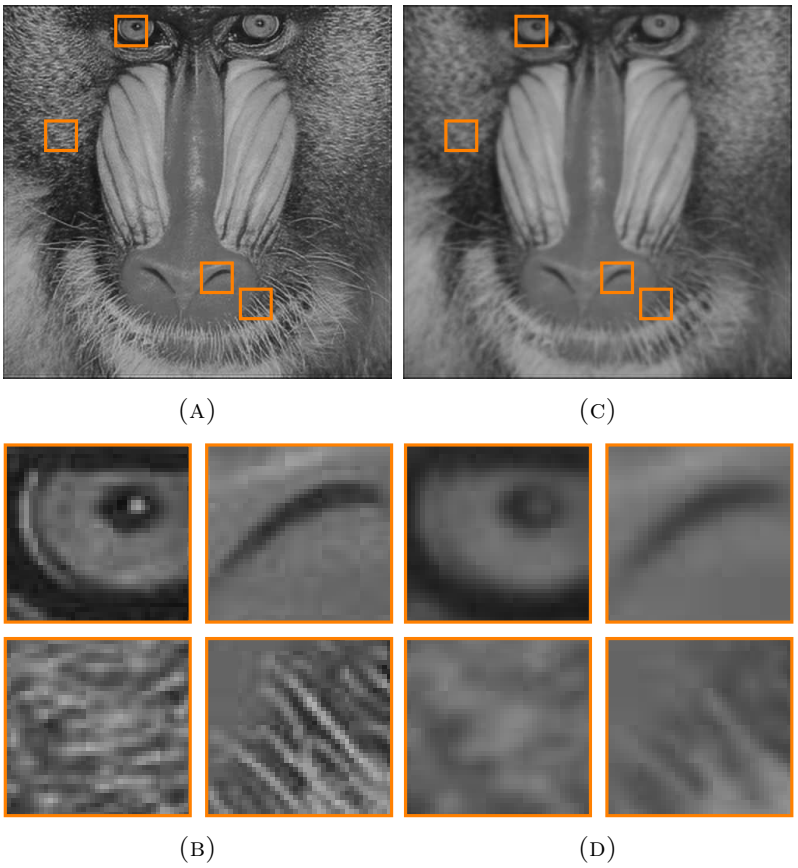


FIGURE 6.5: Test images obtained with the third strategy (inclusive training set), where the two training sets with different blurring kernels ($\sigma = 1.0$ and $\sigma = 2.0$) are fused into one to train the neural network. LR images were generated with $\sigma = 1.0$ for (A, B) and $\sigma = 2.0$ for (C, D).

In the next section, a similar strategy to the inclusive one is adopted, but two main aspects are investigated: a more continuous spanning of the Gaussian kernel space for the observation model, and deeper neural networks to absorb the variability induced by this new observation model.

6.5 Blind and robust Super-Resolution for oriented Gaussian kernels

After this first experimental results, we propose to span a larger and more realistic scope of models by using more than two blurring kernels in the observation model. To compare with state-of-the-art methods, we use the data proposed in [RSRB15].

6.5.1 Problem definition

As in the previous section, the aim of the proposed approach is to recover a SR image $\widetilde{x}$ from a LR image y obtain with the following variable observation model:

$$y = D\bar{B}x \tag{6.3}$$

where x is the HR image, $\bar{B}$ is a variable blurring operator and D a decimation operator that discard every other S pixel for a given scaling factor S. We use Gaussian blur kernels with variable variance and orientation as proposed in [RSRB15]. Figure 6.6 illustrates the generation of different LR images from a single HR sample.

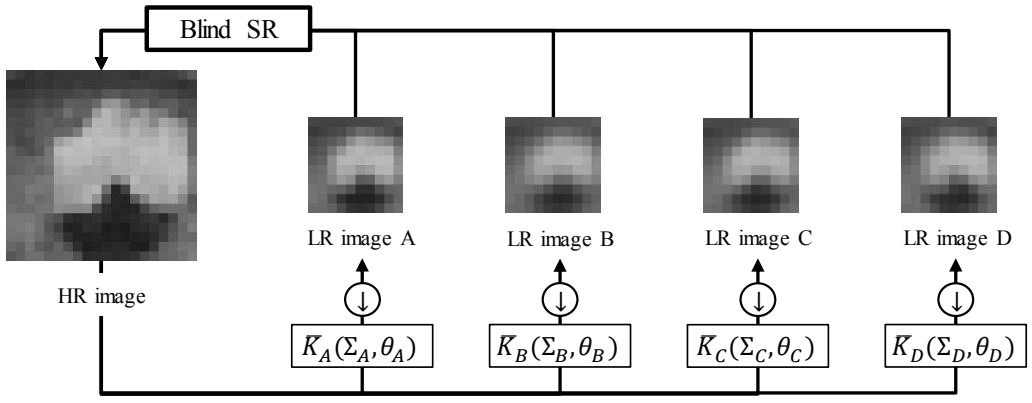


FIGURE 6.6: Generation of several LR images using various blurring kernels from a single HR image and scale factor of 2. Each blurring kernel has different variances and orientations, leading to different LR images.

6.5.2 Proposed approach

We propose to use a single deep CNN to recover a SR image as close as possible to the HR image x. This means we expect it to absorb the different levels of blur and be able to project different visual structures into a HR feature space decorrelated from the applied blur. Figure 6.6 shows an example of such visually different structures that should produce the same HR content. As an input, although other approaches generally use upscaled LR patches, our network takes a LR patch, and performs upsampling at the output layer by using S^2 maps instead of one, rearranged to produce the correct output size. This allows to have a bigger input retina with less 3×3 layers, but requires several output maps. The model can either target the HR patch in graylevel or the high frequencies obtained by difference between the HR patch and upsampled LR patch. Aside from the upsampling approach, the proposed network architecture (Figure 6.7) is very similar to [KLL15a]. It has L layers of 3×3 convolutions, each of which is

fully connected to the previous one; *i.e.* each convolutional kernel has $M \times 3 \times 3 + 1$ parameters including bias, except for the first layer which is directly connected to the input image and the last one that holds $S^2 \times 3 \times 3 + 1$. We use rectified linear units (ReLU) activations after each convolution map. For simplicity, we use zero-padding on borders to keep the feature maps of the same size from the first to the last layer. Using large training patches diminishes the importance of this side effect. We train a CNN with parameters θ and input y to output a full size image $\tilde{x}_i = \Psi(y_i, \theta)$, minimising MSE between the output maps and the target sample.

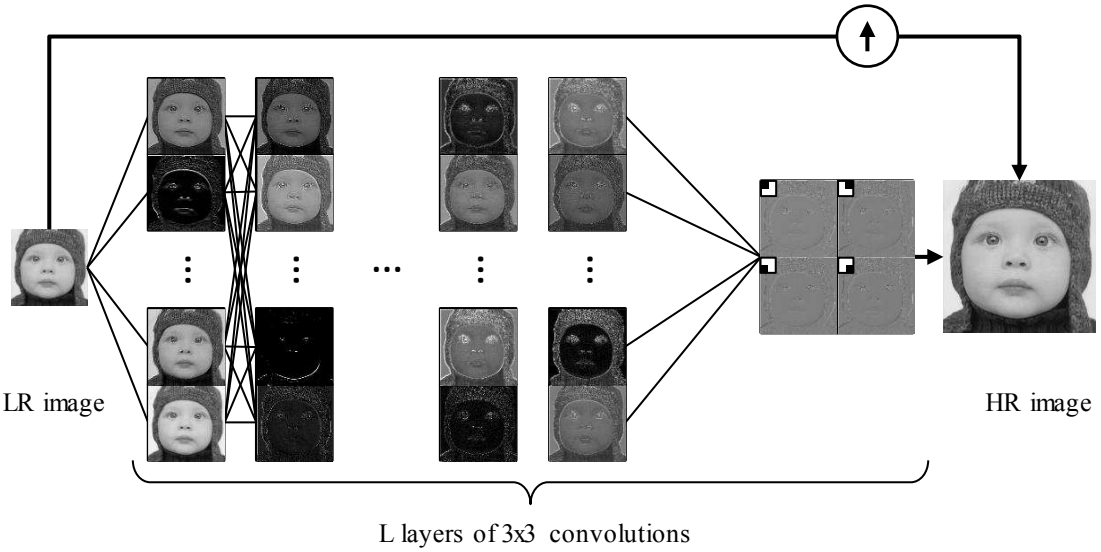


FIGURE 6.7: Proposed Deep CNN for blind SR. The last layer is composed of $S^2 = 4$ maps, rearranged on the HR grid to produce the details missing in the interpolated LR image. Maps dynamic has been modified for visualisation.

In the next section, we provide more details about the experimental architectures and the data used for training.

6.6 Experimental results

6.6.1 Data generation

The data is generated according to [RSRB15] for scale $S = 2$. The LR patches are used as an input *i.e.* without bicubic upsampling. We sample 29,026 pairs of patches from 91 images for each gaussian blurring kernel. These kernels vary in variance and orientation. A total of 58 kernels are used: variances range from 0.75 to 3.0 with a 0.75 step in both dimensions while orientation lies in $[0, \pi]$ with a $\frac{\pi}{8}$ step. This gives a total amount of 1,683,508 example pairs. Input LR patch dimension is 18×18 pixels and 36×36 pixels

for output size. For testing, we use the same procedure on 19 images from *Set5* and *Set14*. Comparative results are presented in subsection 6.6.3.

Given the preliminary study in 6.3, we see that the level of blur is compliant with the limit as variance does not exceeds 3.0 (*i.e.* standard deviation is less than 1.732). 6 kernels were found redundant with the protocol issued by [RSRB15], but we keep the 58 kernels to be able to compare to their results (see Figure 6.8).

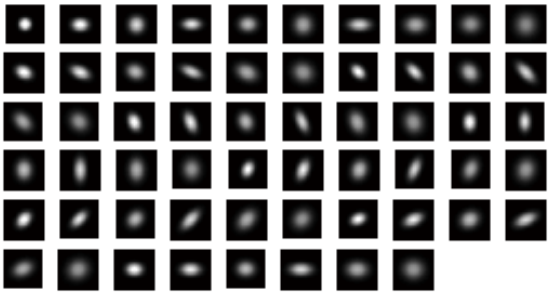


FIGURE 6.8: The different oriented Gaussian kernels used to create the LR images. A total of 58 kernels are used: variances range from 0.75 to 3.0 with a 0.75 step in both dimensions while orientation lies in $[0, \pi]$ with a $\frac{\pi}{8}$ step.

6.6.2 Experiments

Each network is trained from scratch, with a random initialization. We set the global learning rate to 10^{-4}. Although different batch sizes have been tested (4, 16, 128, 256), the reported results were obtained with a pure stochastic gradient descent, which converges faster than using mini-batch and lead to the same order of performance. An hypothesis is that a "mini-batch effect" already takes place, as each target is composed of a 36×36 patch in which each pixel contributes to the loss function. Using mini-batch may therefore lead to more confusing parameter updates during the gradient descent.

We target the high frequencies components instead of the direct graylevel as in the previous chapters 4 and 5. It is also used in [KLL15a]. In Table 6.5, we present the experiments with variations in the number of kernels M and number of layers L. The reported test loss allows to monitor the learning process and select the best models. Test samples are extracted from *Set5*. Although this is not rigorous, it is a common practice in recent SR publications [DLHT14]. In addition, another image set is used for objective performance comparison. We can see that increasing the number of parameters of the proposed model allows to decrease the global MSE. We choose model 7 with 7 layers (comparable to 8) to evaluate on the full test images.

Configuration	L	M	#parameters	Loss
1	4	16	5,380	0.816
2	4	32	19,970	0.766
3	4	64	76,804	0.756
4	4	128	301,060	0.751
5	5	64	113,734	0.739
6	6	64	150,660	0.737
7	7	64	187,588	*0.731*
8	8	64	224,516	**0.730**

TABLE 6.5: 8 configurations with the number of layers L (including the 4-map output layer), the number of kernels per layer M, the total number of parameters and the best obtained test MSE.

	Blind (AB)				Non-blind (CAB)		
	Ours	A+	SRCNN	SRF	A+	SRCNN	SRF
Set5	34.24/**34.52**	33.21	33.58	33.50	33.76	33.92	**34.43**
Set14	**30.82**	30.00	30.27	30.11	30.35	30.50	30.73

TABLE 6.6: PSNR scores (dB) on *Set5* and *Set14*. We report the blind and non-blind results of three experiments of [RSRB15] as a comparison.

6.6.3 Comparison with state-of-the-art example-based SR

We have computed the PSNR for *Set5* and *Set14* for the best obtained model and compared our results to those reported in [RSRB15], using the same protocol, especially cropping 7 pixels at test time to avoid border effects. Results are reported in Table 6.6. We can observe that our approach outperforms the others for the blind set-up. It is also competitive with the non-blind approaches as the mean PSNR is higher than the non-blind A+ and SRCNN methods on *Set5* and the highest for *Set14*, while our approach cannot take advantage of the a priori knowledge of the blurring kernel.

This is a very promising result, as it allows to have similar performance to non-blind approaches which need to know or estimate the precise blurring kernel. For a real-world application, this is an easier solution to implement (a single deep model produces SR images from LR ones) and highly relevant for the variation in LR images.

Resulting images are presented in the next subsection.

6.6.4 Qualitative visual results

SR Images obtained with the selected architecture can be seen in Figures 6.9, 6.10 and 6.11. In each figure, the bicubic interpolation of the LR image and the SR images are displayed. For different Gaussian kernels (smallest variance, largest variance, $-\frac{\pi}{4}$ and $\frac{\pi}{4}$ orientations with largest variance), we can observe various effects. Note that we select the more challenging kernels to outline the behaviour of the trained model for the most extreme cases (very low and very high variance, orientation). First, the produced images are clearly sharper, and edges, textures and objects are better shaped. For instance, hair texture is more rich in Figure 6.10 – (I to P), and letters are more readable in Figure 6.11 – (I to P).

For the kernel with smallest variance, the networks does not produce overshooting artefacts like in 6.4, but still has a strong response on edges. For instance, repetitive artefacts appears in Figure 6.10 – J, exaggerating the skin texture. For the largest variance kernel, the network recovers sharp edges and texture details as can be seen on Figures 6.10 and 6.11 (last column of each figure: G,H,P,O)

With oriented kernels, the SR images are sharper but also exhibit oriented artefacts that the deep model does not compensate. This is particularly visible on strong edges such as close-ups N and P in Figure 6.9: when the edge has the same orientation as the oriented kernel, it is accurately reconstructed (N, bottom left), but not when direction are perpendicular (P, bottom left). A possible explanation of these remaining artefacts is that the CNN often compromises by averaging the possible orientations, ending up in an efficient but non-oriented high-frequency compensation.

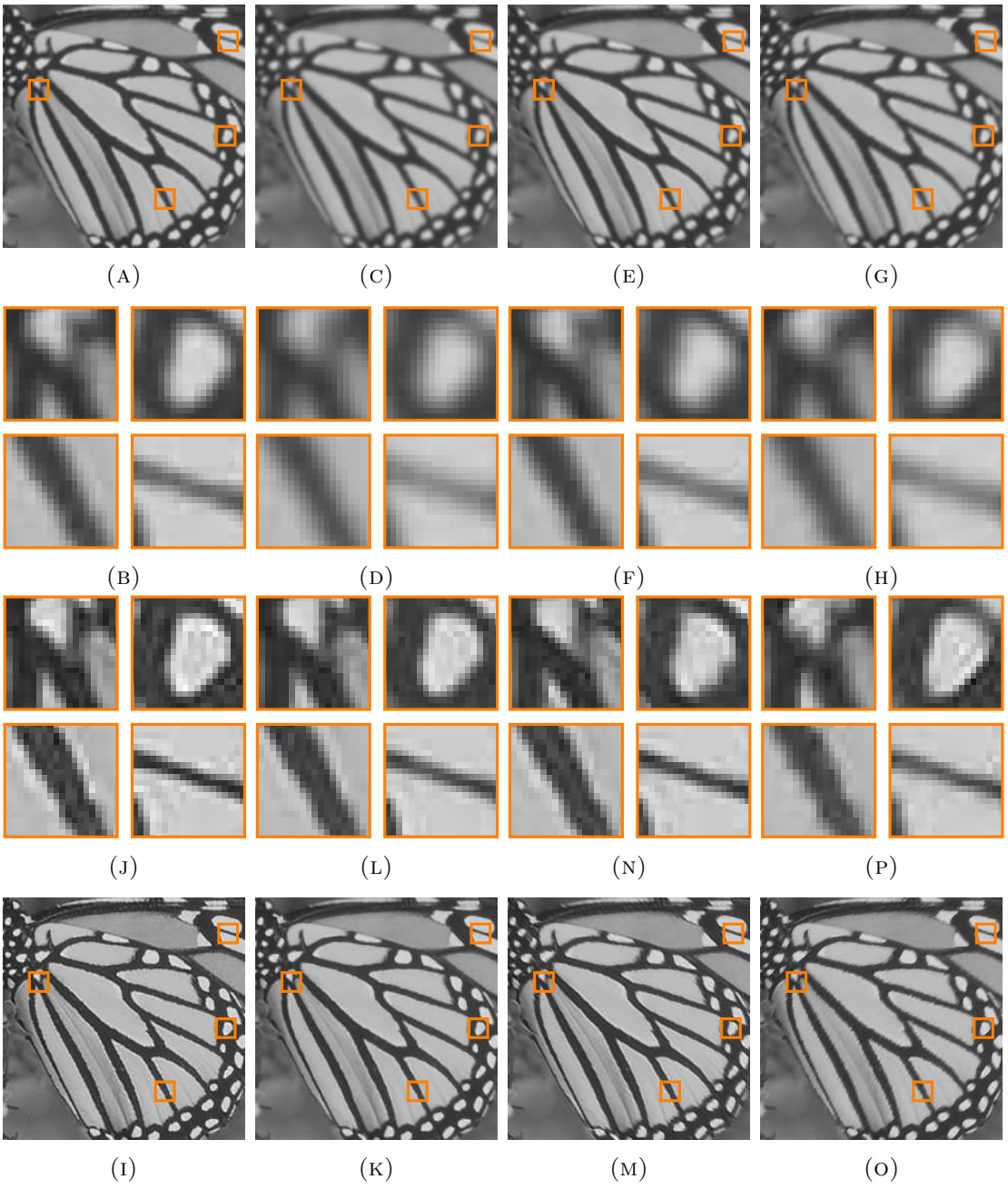


FIGURE 6.9: Results for the *Butterfly* test image, rich in edges in all orientations, with four different blurring kernels used in the observation model. Compared with a bicubic interpolation (A – H), the blur artefacts are well removed in the SR images (I – P). Some overshooting is present for the directions with small variance (I-J, M-N and OP).

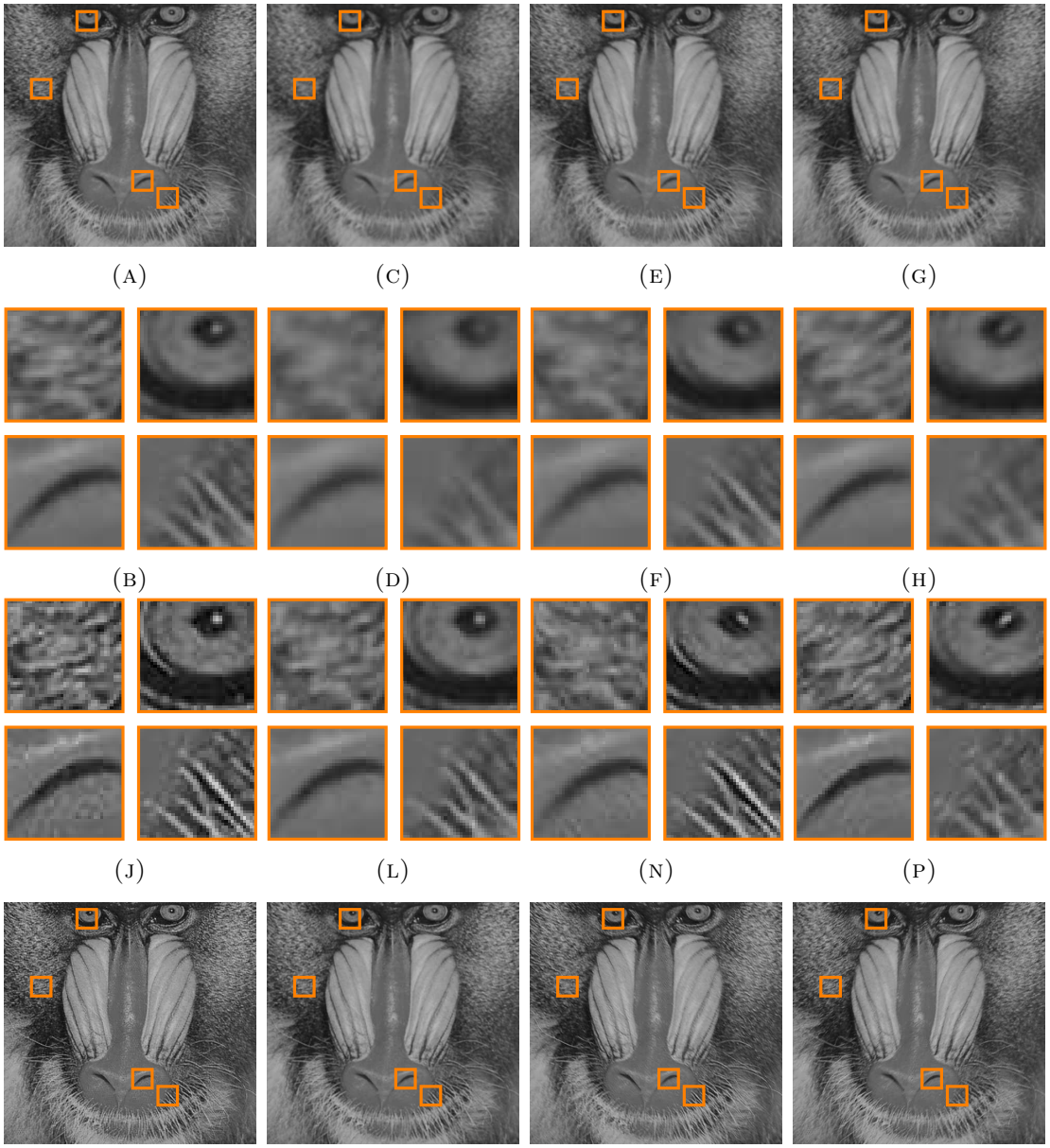


FIGURE 6.10: Results for the *Mandrilla* test image. Compared with a bicubic interpolation (A – H), the hair regions in the SR images (I – P) are more detailed, as well as the eye glow. The skin texture of the nose is slightly exaggerated in close-up J, bottom left. Strongly oriented kernels (E – H for bicubic and M – P for SR) also present oriented artefact that the network cannot compensate.

FIGURE 6.11: Results for the *Powerpoint* test image. Compared with a bicubic interpolation (A – H), the obtained SR images (I – P) allow a better readability of the textual content, and sharp edges. Some blur artefacts are still present, particularly on oriented kernels (M,N,O,P).

6.7 Conclusion

We have presented a blind approach to Super-Resolution using a Deep Convolutional Neural Network architecture. The network is trained with LR and HR image pairs where LR images are produced with different blurring kernels. Although shallow networks perform well for the non-blind set-up [DLHT14], the experiments show that by using more parameters and deeper neural models than in previous work, we can improve the robustness of CNN-based models for blind SR.

In particular, results show that the proposed method achieves better reconstruction performance than the previous results reported on the same dataset in [RSRB15], for the blind set-up and also the non-blind one, where the observation model is fed to the reported SR approaches. By being robust to different observation models, the proposed method can be deployed for real-world. For instance, it can be used to upscale images taken with different devices that have different sensors and lenses that influence the real observation model. It can also cope with small motion thanks to the improved robustness on oriented kernels.

Chapter 7

Conclusion

Contents

In this last chapter, a conclusion on the work presented throughout this thesis is given in section 7.1: the three contributions on text, face and blind SISR are summarised. Then, the limitations of the proposed approaches are discussed in section 7.2: for each one, the gaps that would allow to address the largest scope of images in the best conditions are analysed. Finally, section 7.3 is composed of two parts. In section 7.3.1, comments on trends and perspective in the SR field are given. Then, consistent with the course of this document, a reflection on possible extension of this work is proposed in section 7.3.2. With roots in the latest advances in SR, the proposed track aims to replace the hand-crafted knowledge-based prior with an automatic one, coming from a recognition system that inject its own prior on domain-specific high-resolution images. A list of associated publications is provided in section 7.4.

7.1 Summary of the contributions

This work has focused on SISR methods designed to improve not only the perceptual quality of the images, but also the performances of automatic recognition systems.

The first core contribution of this work is a SR method based on artificial neural networks (Multi-Layer Perceptrons or Convolutional Neural Networks) for text image resolution enhancement. This example-based approach allows to improve both the image reconstruction quality and the accuracy of an OCR system, on document images and text images extracted from TV streams. For document images, the proposed method allows to improve reconstruction by $+7.23dB$ in PSNR and accuracy by +7.85 points compared with a standard bicubic interpolation of the LR images, setting new state of the art results on the *ULR-textsisr-2013a* dataset. An advanced analysis of the proposed approach shows how the neural models allow to automatically learn relevant features and mappings in an end-to-end fashion. The results indicate that CNN models reach better performance by incorporating hierarchical, highly non-linear LR feature extraction and mapping. To address multimedia content, a special dataset on text images from TV streams have been publicly released for the organisation of the first international competition of text image super-resolution for the ICDAR2015 conference. Results show that learning-based methods are highly relevant for this task. In particular, the best results obtained on this new dataset make use of CNN ensembles with performance close to the one obtained on the original HR image set. For this first study on text, the adaptation to the context is made solely by adapting the training data to the task, to obtain dedicated systems.

To tackle the equivalent problem on unconstrained face SR, a second approach is presented. It capitalises on the first proposed model and extends it in a two step fashion. The first step improves the resolution of the global image via a local model, using a CNN trained on large-scale facial dataset. It allows to improve reconstruction and recognition scores over the low-resolution images, respectively by +3.44dB (compared with bicubic interpolation) and +6.91 points compared with the original LR images and +2.70 over bicubic interpolation. The second step consists in detecting facial components (eyes, nose, mouth) and improving them using dedicated models, that recover severe degraded components. Based on autoencoder architectures, they transform the input components into a non-negative, part-based representation from which the final SR components are synthesised. The second step allows to further improve the recognition performance reaching +8.15 points (+3.94 compared with bicubic interpolation), even though the pixel-wise measures are slightly degraded.

The third contribution is centred on real-world aspects and the relevance of the proposed neural based approaches in such contexts. Indeed, real LR images may be obtained through various devices and situations. This is modelled via a variation of the blurring kernel of the observation model. To tackle these variations, a deep CNN is learned in a blind fashion (*i.e.* without knowledge of the blurring kernel used in the observation model). It is demonstrated that, contrary to the limits observed with a single observation model, having deeper architectures improve performance over the whole dataset, even surpassing kernel-aware methods proposed in the literature. The images obtained on natural dataset do not exhibit overshooting artefacts as observed in preliminary study with shallow models. The obtained results indicate promising real-world application of such approach, due to the improved robustness and the accurate reconstructed SR images.

7.2 Limitations of the proposed approaches

The methods described in the first contribution in chapter 4 rely on the choice of the data. This provides specialised systems that are efficient when the conditions imposed by the training examples are met. However, as reported in Figure 4.12 (section 4.4, chapter 4), the best obtained models may still face difficult and ambiguous local contexts in the LR image, where they produce irrelevant or blurry predictions. Also, the models trained on a specific kind of data (*e.g. document images*) can be required to function on new type of images (*e.g.* text extracted from TV streams), and the performance increase observed in the experimental set-up might be less noticeable if those new cases are too far from the training data. Moreover, the data is generated following a fixed observation model, which may not be compliant with real-world observations, subject to more variation (moving objects, blurry or noisy conditions). Addressing such images would also require more robust models, either by selecting adapted data or by incorporating mechanisms such as noise reduction or outlier detection for uncommon samples.

For face SISR (chapter 5), the proposed method rely on two steps and is dependant on external tools such as the facial landmarks detector. Having a unified network for both the generic and the specific steps while conserving the strength of the approach could be beneficial. Also, even if the reconstructed components by the second step are more realistic than after the first step, they lack of high-frequency content to hallucinate credible high-resolution regions. As mentioned at the end of the chapter, a finer selection of the training images could be beneficial.

In chapter 6, although the proposed approach clearly improves the PSNR in the blind set-up, strongly oriented kernels produce artefacts that are difficult to compensate. An

alternative would be to predict the blurring kernel associated to a non-blind set-up that would use the predicted kernel to compensate the right artefacts, such as those in [RSRB15]. Such approaches have been recently proposed for blind deblurring [SCXP15, Cha16] and seem promising. Another choice could be to force the neural network to *explicit* its estimation of the underlying blurring kernel (*e.g.* by predicting it), in order to have a control on whether the network interprets correctly the amount and orientation of the blur contained in the LR image or not.

Another track of investigation is to force the abstraction in the CNN, to make it less dependent on the input image and its potential artefacts. Even if more robust, the proposed model is composed of fine grain 3×3 convolutions that allow to carry localized errors through the different layers until the output one. According to recent work[BSL16], blind SR could profit from spatial abstraction via pooling layers. It might constrain the encoding process of the CNN to extract meaningful internal representation independently from the spatial inconvenience of the different blurring effects.

7.3 Future works

7.3.1 Perspectives for Super-Resolution

Recent successful approaches in automatic image synthesis suggest that Generative Adversarial Networks (GANs) achieve unprecedented results. The adversarial training takes place using two neural networks (one Generator and one Discriminator) that compete with each other: while the generator learns to fool the discriminator better, the latter benefits from more and more challenging generated examples. This training procedure being paired with hierarchical representations, GANs produce relevant images on both local (textures, image grain) and global scales (object parts,). Results on face hallucination and natural image SR have been demonstrating how relevant these model are. However, two points could be worth exploring. First, although the upscaling factors are generally high, it is also unclear how such approaches are dependant on the observation model, as they often use a single LR observation model (bicubic downsampling). Second, it would be interesting to design an approach such as [LTH+16] based on different criterion than photo realism (achieved via SR/HR image discrimination). A proposal is presented in the next subsection to address this challenge.

Another challenge that was not addressed in this thesis involve multiple image SR using neural networks. Indeed, some very low resolution content can only be reconstructed into higher resolution images by gathering the data split in several images. Also, using SISR methods to upscale videos is inefficient as the redundant information between frames on

which compression methods are based is likely to have the same appearance in the HR frames. Recent work have been proposed to tackle this problem using recurrent networks or 3D convolutions in CNN [HWW15]. They manage to reduce the aliasing artefact of HR images. However, the gain in PSNR and image quality does not set impressive gaps (less than 1dB) compared with SISR methods, and the methods tend to use synthetic data and MSE training, rather than taking profit of the technological shift proposed by methods such as GANs.

7.3.2 A preliminary study on Task-Guided Super-Resolution

Recent trends in example-based Super-Resolution and image enhancement include perception-based SR, where the optimisation criterion of a learning system is not based on a pixel-wise cost with additive priors on the model parameters, but rather *perception* losses where an intelligent system is able to score the quality of the obtained image, and moreover give a feedback to contribute to the general optimisation of the learning system. Although the trend is quite recent, the obtained results are really impressive and encouraging. Figure 7.1 depicts several results obtained by [JAFF16] and [LTH+16]. Although their approach diverge in philosophy – [JAFF16] uses a style-transfer approach while [LTH+16] explicit the perceptual mechanism using adversarial nets that compete on a "SR or real image ?" task – they both aim to bring satisfactory image statistics that comply with the realism that HVS is used to when confronted to the real world. Throughout this work we have focus on a interesting mission: to perform SR on specific

(A) Bicubic (×4) (B) Result from [JAFF16] (C) Result from [LTH+16] (D) Original HR

FIGURE 7.1: Results from recent works, involving an automatic perception of the produced SR images to make an image quality feedback available to the training process.

images that aim to be processed by recognition engines, texts and facial images. For text images, we have shown in chapter 4 that SR can help to boost the performance of recognition engines while giving pleasant visual results, simply by adapting the data

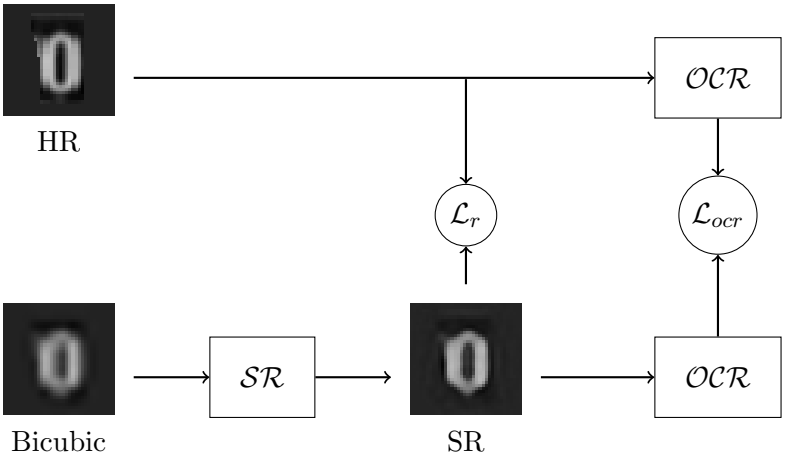


FIGURE 7.2: Proposed approach for Task-Guided Super-Resolution. The $\mathcal{SR}$ network is trained with respect to two loss functions: $\mathcal{L}_r$ which is the usual regression loss and the $\mathcal{L}_{ocr}$ one, which is given by the difference between internal $\mathcal{OCR}$ network activations for SR and HR images.

to the nature of images. We have seen that even in one type of images – text – a certain variety can occur and that specialisation can help (see chapter 4, section 4.4 on document image and section 4.5 on TV content). In chapter 5 we observed that incorporating domain knowledge can also help recognition system, as we know the preference of such systems. The natural path to take, in convergence with the trends we mentioned, would be to connect the domain knowledge not by incorporating hand-crafted tricks but using the preferences of the recognition engine itself. Indeed, the important low-level and high-level features that a given engine extracts to perform its classification task may provide feedback. If the derivative with respect to the SR engine parameters are tractable through this recognition system, it is possible to incorporate this information in the learning process.

7.3.2.1 Proposed track

In order to benefit from an automated feedback, we present the idea of a Task-Guided Super-Resolution (TGSR) approach. Instead of updating the parameters of a network using exclusively the residual between the SR output and the original HR image, the images produced by the SR network are presented to a second static network, trained for character recognition as described in [SG07]. The penultimate layer of this deep neural network is selected to give a high dimensional vector of neural activations. The activation vector obtained with the SR image can be compared with the one obtained with the original HR image, making a new loss function available. This vector is still correlated with the input image as different images give different activations. However,

the network projects this input stimuli into an abstract space where images are discriminable depending on their character class. Therefore, a large projection error implies a higher likelihood to end up with a misclassified image. The squared difference between the two feature vectors defines the following OCR Loss:

$$\begin{aligned}\mathcal{L}_{ocr} &= \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_{ocr}(x_k, y_k) \\ &= \frac{1}{K}\sum_{k=1}^{K}\Big[\mathcal{OCR}\big(\mathcal{SR}\left(x_k\right)\big) - \mathcal{OCR}\left(y_k\right)\Big]^2\end{aligned} \tag{7.1}$$

where $\mathcal{OCR}$ represent the OCR non-linear function, $\mathcal{SR}$ the one of the SR network being trained, K is the number of training samples. To guide the network to produce relevant images, a normal reconstruction loss $\mathcal{L}_r$ is used

$$\begin{aligned}\mathcal{L}_r &= \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_r(x_k, y_k) \\ &= \frac{1}{K}\sum_{k=1}^{K}\Big[\mathcal{SR}\left(x_k\right) - y_k\Big]^2\end{aligned} \tag{7.2}$$

The overall loss is:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{OCR} + \beta \cdot \mathcal{L}_{sr} \tag{7.3}$$

where α and β control the influence of each loss on the training process.

The aim is to minimise this loss over a training set, by tuning the parameters of the $\mathcal{SR}$ net only, as the $\mathcal{OCR}$ network is static and only provides the loss and the useful gradients. Using stochastic gradient descent, for a given training sample x_k, each parameter θ_i of the $\mathcal{SR}$ network is update following the gradient:

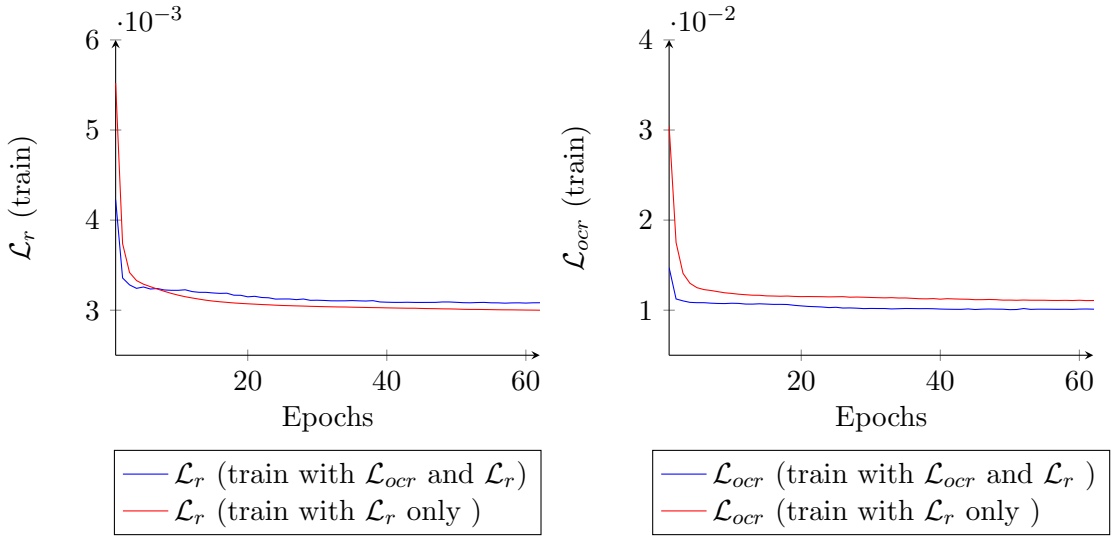
$$\theta_i^t = \theta_i^{t-1} + \eta\frac{\partial\mathcal{L}(x_k, y_k)}{\partial\theta_i} \tag{7.4}$$

Note that even though the $\mathcal{OCR}$ network uses the *global* SR image to classify it, the SR network only performs *local* processing, using a sliding 13×13 retina. Therefore, all it can learn is a *local* image prior and not an *global* transformation of the image into a more recognisable object.

7.3.2.2 Preliminary experimental results

We use a training set composed of 15,000 segmented character images from various literature datasets. A test set composed of another 15,000 images is used to evaluate

the model. The downsampling factor is $s = 3$, to sufficiently degrade the original images. For simplicity, α and β are set to 1, meaning that each loss contributes equally in the parameter update. Setting α to zero gives divergent behaviour (pure task guided SR), as there is no regularisation from the $\mathcal{L}_r$ loss. The same architecture is also trained using a reconstruction criterion only, *i.e.* minimising the $\mathcal{L}_r$ loss. Figure 7.3 an interesting behaviour during training, as it relates to previous observations made in chapter 5 on the two step approach. Compared with the training using $\mathcal{L}_r$ only, the use of multiple loss $\mathcal{L}_r$ and $\mathcal{L}_{ocr}$ gives a higher average loss for reconstruction, but a lower average loss for the OCR.



(A) Evolution of the reconstruction loss $\mathcal{L}_r$ during training, using the proposed multiple criteria or only the reconstruction one.

(B) Evolution of the reconstruction loss $\mathcal{L}_r$ during training, using the proposed multiple criteria or only the reconstruction one.

FIGURE 7.3: Training loss monitoring with or without the feedback from the $\mathcal{OCR}$ network. A trade-off is observed between a better reconstruction (lower $\mathcal{L}_r$) and a more accurate OCR representation (lower $\mathcal{L}_{ocr}$).

The OCR results obtained on the test set are reported on Table 7.1. We can see an increase in the OCR accuracy, the same way face verification increased when injecting knowledge about the recognition engine. However, while this knowledge was injected using a hand-crafted architecture, it is now brought directly by the automatic recognition system itself via backpropagation. Slight variations are observed between the resulting images of the two SR approaches. As depicted in Figure 7.4, the images resulting from the TGSR approach exhibit locally boosted edges, improving the global acuity and sharpness. This corresponds to the local prior assumption, and gives the first insight on how gathering automatic feedback from a recognition system may help to improve task-specific SR system, without incorporating hand-crafted architectures or assumptions on the recognition engine. These results could be improved in several aspects:

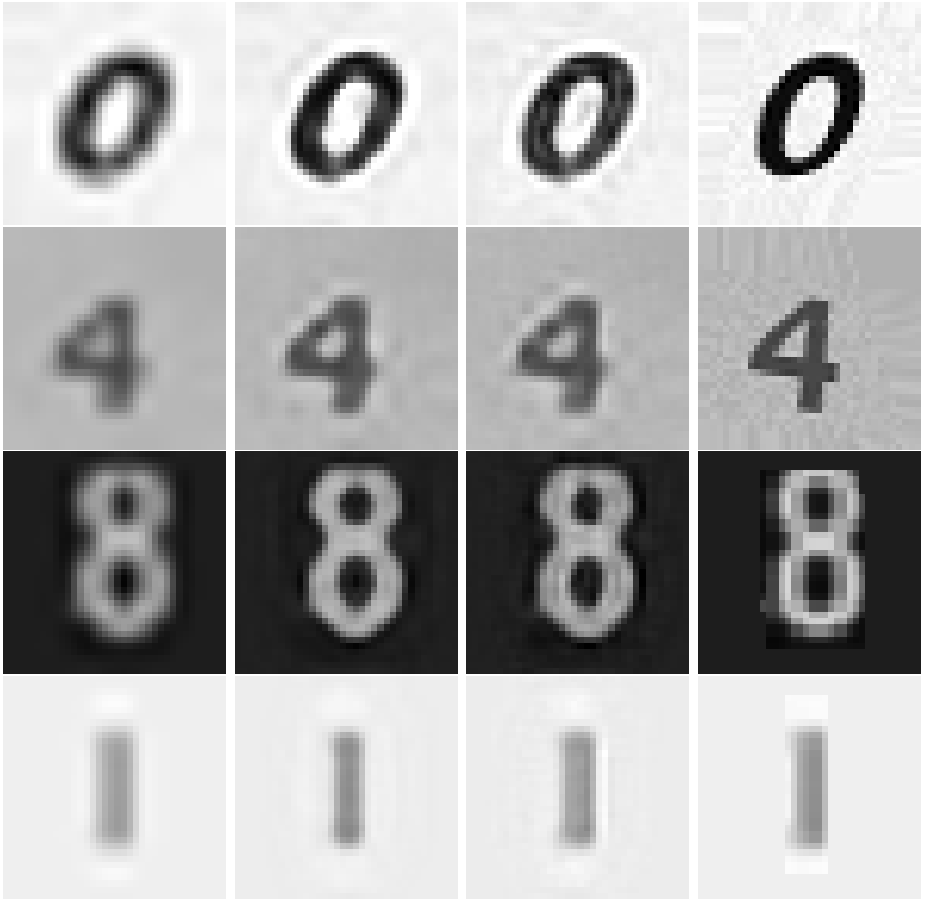
TABLE 7.1: OCR results obtained on the test set.

Images	OCR character accuracy
Bicubic	79.253
SR ($\mathcal{L}_r$)	83.680
TGSR ($\mathcal{L}_r + \mathcal{L}_{ocr}$)	**84.087**
HR	92.260

1. a larger dataset with high-resolution character images could be used as the one used for training contain some images with poor quality.

2. a fine-tuning of the update parameters could allow a more accurate prior to be learned by the $\mathcal{SR}$ network: learning rates, optimisation schemes, dynamic balance between the two loss functions.

3. the ultimate end-to-end approach would be to use the classification label instead of the penultimate layer.



(A) Bicubic (×4) (B) SR (C) TGSR (D) HR

FIGURE 7.4: Comparative results for the proposed approach. The SR method produces cleaner images while the TGSR one generate overshoot artefacts that increase acuity, and improves recognition performance of the OCR engine.

7.4 List of publications

International Conferences

- Clément Peyrard, Franck Mamalet, and Christophe Garcia. "A Comparison between Multi-Layer Perceptrons and Convolutional Neural Networks for Text Image Super-Resolution." In International Conference on Computer Vision Theory and Applications, pp. 84-91. 2015.
- Clément Peyrard, Moez Baccouche, Franck Mamalet, and Christophe Garcia. "ICDAR2015 competition on Text Image Super-Resolution." In International Conference on Document Analysis and Recognition, pp. 1201-1205., 2015.
- Guillaume Berger, Clément Peyrard, and Moez Baccouche. "Boosting face recognition via neural Super-Resolution." In European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2016
- Clément Peyrard, Moez Baccouche, and Christophe Garcia. "Blind Super-Resolution with Deep Convolutional Neural Networks." In International Conference on Artificial Neural Networks, pp. 161-169. 2016.

National Workshop

- Clément Peyrard, Moez Baccouche, and Christophe Garcia. "Deep Learning methods for Image Super-Resolution" In Apprentissage Profond (Deep Learning), GdR ISIS, Paris, 2016.

Bibliography

[AEB06] Michal Aharon, Michael Elad, and Alfred Bruckstein. *rmk*-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322, 2006.

[AGK95] F. Ahmed, S.C. Gustafson, and M.a. Karim. High-fidelity image interpolation using radial basis function neural networks. *IEEE National Aerospace and Electronics Conference*, 2:588–592, 1995.

[AW96] Jan Allebach and Ping Wah Wong. Edge-directed interpolation. In *International Conference on Image Processing*, volume 3, pages 707–710. IEEE, 1996.

[BCM05] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 60–65. IEEE, 2005.

[BJKK13] Peter N Belhumeur, David W Jacobs, David J Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2930–2940, 2013.

[BK00] Simon Baker and Takeo Kanade. Hallucinating faces. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 83–88. IEEE, 2000.

[BK02] Simon Baker and Takeo Kanade. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1167–1183, sep 2002.

[BM13] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

[BMW+11] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. Sequential deep learning for human action recognition. In *International Workshop on Human Behavior Understanding*, pages 29–39. Springer, 2011.

[BRGM12] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *British Machine Vision Conference (BMVC)*, 2012.

[BSFG09] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: a randomized correspondence algorithm for structural image editing. In *ACM Transactions on Graphics*, 2009.

[BSL16] Joan Bruna, Pablo Sprechmann, and Yann LeCun. Super-resolution with deep convolutional sufficient statistics. *International Conference on Learning Representations*, 2016.

[BT09] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[CCS+14] Zhen Cui, Hong Chang, Shiguang Shan, Bineng Zhong, and Xilin Chen. Deep network cascade for image super-resolution. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014.

[CCvBS15] Xu Chen, Anustup Choudhury, Peter van Beek, and Andrew Segall. Facial video super resolution using semantic exemplar components. In *IEEE International Conference on Image Processing*, pages 1314–1318. IEEE, 2015.

[Cha16] Ayan Chakrabarti. A neural approach to blind motion deblurring. In *European Conference on Computer Vision*, pages 221–235. Springer, 2016.

[CS14] Anustup Choudhury and Andrew Segall. Channeling mr. potato head-face super-resolution using semantic components. In *IEEE Southwest Symposium on Image Analysis and Interpretation*, pages 157–160. IEEE, 2014.

[CYX04] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-resolution through neighbor embedding. In *Computer Vision and Pattern Recognition*, volume 1, pages I–I. IEEE, 2004.

[CZ00] David Capel and Andrew Zisserman. Super-resolution enhancement of text image sequences. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 1, pages 600–605. IEEE, 2000.

[CZ01] David Capel and Andrew Zisserman. Super-resolution from multiple views using learnt image models. In *Computer Vision and Pattern Recognition*, volume 2, pages II–627. IEEE, 2001.

[DBBE+09] Albertus C Den Brinker, Jeroen Breebaart, Per Ekstrand, Jonas Engdegård, Fredrik Henn, Kristofer Kjörling, Werner Oomen, and Heiko Purnhagen. An overview of the coding standard mpeg-4 audio amendments 1 and 2: He-aac, ssc, and he-aac v2. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009(1):468971, 2009.

[DDDM04] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on pure and applied mathematics*, 57(11):1413–1457, 2004.

[DG07] S. Duffner and C. Garcia. Face recognition using non-linear image reconstruction. In *International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, pages 459–464, London, UK, September 2007.

[DG08] Manolis Delakis and Christophe Garcia. text detection with convolutional neural networks. In *VISAPP*, 2008.

[DH00] Carlos A Dávila and BR Hunt. Superresolution of binary images with a nonlinear interpolative neural network. *Applied Optics*, 39:2291–2299, 2000.

[DLHT14] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*, 2014.

[DLHT16] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2016.

[DM05] Katherine Donaldson and Gregory K Myers. Bayesian super-resolution of text in videowith a text-specific bimodal prior. *International Journal of Document Analysis and Recognition (IJDAR)*, 7(2-3):159–167, 2005.

[DZD+15] Chao Dong, Ximei Zhu, Yubin Deng, Chen Change Loy, and Yu Qiao. Boosting optical character recognition: A super-resolution approach. *arXiv preprint arXiv:1506.02211*, 2015.

[DZSW11] Weisheng Dong, Lei Zhang, Guangming Shi, and Xiaolin Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing*, 20(7):1838–1857, 2011.

[EGA+13] Netalee Efrat, Daniel Glasner, Alexander Apartsin, Boaz Nadler, and Anat Levin. Accurate blur models vs. image priors in single image super-resolution. In *IEEE International Conference on Computer Vision*, pages 2832–2839, 2013.

[EGMS14] Khaoula Elagouni, Christophe Garcia, Franck Mamalet, and Pascale Sébillot. Text recognition in multimedia documents: a study of two neural-based ocrs using and avoiding character segmentation. *International Journal on Document Analysis and Recognition (IJDAR)*, 17(1):19–31, 2014.

[EGS11] Khaoula Elagouni, Christophe Garcia, and Pascale Sébillot. A comprehensive neural-based approach for text recognition in videos using natural language processing. In *ACM International Conference on Multimedia Retrieval*, page 23. ACM, 2011.

[Ela13] Khaoula Elagouni. *Combining neural-based approaches and linguistic knowledge for text recognition in multimedia documents.* PhD thesis, Rennes, INSA, 2013.

[Eri17] Ericsson. Ericsson mobility report. Technical report, Ericsson, June 2017.

[EV07] Mehran Ebrahimi and Edward R Vrscay. Solving the inverse problem of image zooming using "self-examples". In *International Conference Image Analysis and Recognition*, pages 117–130. Springer, 2007.

[Fat07] Raanan Fattal. Image upsampling via imposed edge statistics. *ACM Transactions on Graphics*, 26(3):95, jul 2007.

[FF11] Gilad Freedman and Raanan Fattal. Image and video upscaling from local self-examples. *ACM Transactions on Graphics*, 30(2):1–11, apr 2011.

[FJP02] William T Freeman, Thouis R Jones, and Egon C Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, 22(2):56–65, 2002.

[FKMI09] Wataru Fukuda, Atsunori Kanemura, Shin-ich Maeda, and Shin Ishii. Superresolution from occluded scenes. In *International Conference on Neural Information Processing*, pages 19–27. Springer, 2009.

[FPC00] William T Freeman, Egon C Pasztor, and Owen T Carmichael. Learning low-level vision. *International journal of computer vision*, 40(1):25–47, 2000.

[GAF+87] John D Gould, Lizette Alfaro, Rich Finn, Brian Haupt, and Angela Minuto. Reading from crt displays can be as fast as reading from paper. *Human factors*, 29(5):497–517, 1987.

[GBA+03] Bahadir K Gunturk, Aziz Umit Batur, Yucel Altunbasak, Monson H Hayes, and Russell M Mersereau. Eigenface-domain super-resolution for face recognition. *IEEE transactions on image processing*, 12(5):597–606, 2003.

[GBI09] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. *IEEE 12th International Conference on Computer Vision*, pages 349–356, sep 2009.

[GD04] Christophe Garcia and Manolis Delakis. Convolutional face finder: A neural architecture for fast and robust face detection. *IEEE Transactions on pattern analysis and machine intelligence*, 26(11):1408–1423, 2004.

[GGY13] Junbin Gao, Yi Guo, and Ming Yin. Restricted boltzmann machine approach to couple dictionary training for image super-resolution. In *IEEE International Conference on Image Processing*, pages 499–503. IEEE, 2013.

[GL10] Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *International Conference on Machine Learning*, pages 399–406, 2010.

[GO04] Tomomasa Gotoh and Masatoshi Okutomi. Direct super-resolution and registration using raw cfa images. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–600. IEEE, 2004.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[GSL00] Jinwook Go, Kwanghoon Sohn, and Chulhee Lee. Interpolation using neural networks for digital still cameras. *IEEE Transactions on Consumer Electronics*, 46(3):610–616, 2000.

[GZTL12] Xinbo Gao, Kaibing Zhang, Dacheng Tao, and Xuelong Li. Image super-resolution with sparse neighbor embedding. *IEEE Transactions on Image Processing*, 21(7):3194–3205, 2012.

[HBA97] Russell C Hardie, Kenneth J Barnard, and Ernest E Armstrong. Joint map registration and high-resolution image estimation using a sequence of undersampled images. *IEEE Transactions on Image Processing*, 6(12):1621–1633, 1997.

[HRBLM07] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[HS11] He He and Wan-Chi Siu. Single image super-resolution using gaussian process regression. In *Computer Vision and Pattern Recognition*, pages 449–456. IEEE, 2011.

[HSA15] Jia-bin Huang, Abhishek Singh, and Narendra Ahuja. Single Image Super-resolution from Transformed Self-Exemplars. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197—-5206, 2015.

[HSCG10] Hu Han, Shiguang Shan, Xilin Chen, and Wen Gao. Gray-scale super-resolution for face recognition from low gray-scale resolution face images. In *IEEE International Conference on Image Processing*, pages 2825–2828. IEEE, 2010.

[Hua84] TS Huang. Multi-frame image restoration and registration. *Advances in computer vision and Image Processing*, 1:317–339, 1984.

[HWW15] Yan Huang, Wei Wang, and Liang Wang. Bidirectional recurrent convolutional networks for multi-frame super-resolution. In *Advances in Neural Information Processing Systems*, pages 235–243, 2015.

[IP91] M Irani and S Peleg. Improving resolution by image registration. *Graphical models and image processing*, 1991.

[JAFF16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *European Conference on Computer Vision*, 2016.

[JG08] Kui Jia and Shaogang Gong. Generalized face super-resolution. *IEEE Transactions on Image Processing*, 17(6):873–886, 2008.

[JHH+13] Junjun Jiang, Ruimin Hu, Zhen Han, Zhongyuan Wang, Tao Lu, and Jun Chen. Locality-constraint iterative neighbor embedding for face hallucination. In *IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2013.

[JHWH14] Junjun Jiang, Ruimin Hu, Zhongyuan Wang, and Zhen Han. Face super-resolution via multilayer locality-constrained iterative neighbor embedding and intermediate dictionary learning. *IEEE Transactions on Image Processing*, 23(10):4220–4231, 2014.

[JMR+07] Viren Jain, Joseph F Murray, Fabian Roth, Srinivas Turaga, Valentin Zhigulin, Kevin L Briggman, Moritz N Helmstaedter, Winfried Denk, and H Sebastian Seung. Supervised learning of image restoration with convolutional networks. In *International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[KBBN09] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 365–372. IEEE, 2009.

[KGBN+15] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 1156–1160. IEEE, 2015.

[KLL15a] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. *arXiv preprint arXiv:1511.04587*, 2015.

[KLL15b] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. *arXiv preprint arXiv:1511.04491*, 2015.

[KSB+10] Koray Kavukcuoglu, Pierre Sermanet, Y-Lan Boureau, Karol Gregor, Michaël Mathieu, and Yann L Cun. Learning convolutional feature hierarchies for visual recognition. In *Advances in neural information processing systems*, pages 1090–1098, 2010.

[Kum03] N. Kumar. A novel neural-network-based image resolution enhancement. *IEEE International Conference on Fuzzy Systems*, 2:1428–1433, 2003.

[LBBH98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LBRN06] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y Ng. Efficient sparse coding algorithms. In *Advances in neural information processing systems*, pages 801–808, 2006.

[LFP+07] Chin-Teng Lin, Kang-Wei Fan, Her-Chang Pu, Shih-Mao Lu, and Sheng-Fu Liang. An HVS-Directed Neural-Network-Based Image Resolution Enhancement Scheme for Image Resizing. *IEEE Transactions on Fuzzy Systems*, 15(4):605–615, aug 2007.

[Lie03] Rainer Lienhart. Video ocr: A survey and practitioner's guide. In *Video mining*, pages 155–183. Springer, 2003.

[LKD99] Huiping Li, Omid E Kia, and David S Doermann. Text enhancement in digital video. In *Electronic Imaging'99*, pages 2–9. International Society for Optics and Photonics, 1999.

[LO01] X Li and M T Orchard. New edge-directed interpolation. *IEEE transactions on image processing*, 10(10):1521–7, jan 2001.

[LP93] Seong Won Lee and Joon Ki Paik. Image interpolation using adaptive fast b-spline filtering. In *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, volume 5, pages 177–180. IEEE, 1993.

[LP07] H Luong and Wilfried Philips. Non-local text image reconstruction. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 1, pages 546–550. IEEE, 2007.

[LS99] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[LS01] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[LS11] Ce Liu and Deqing Sun. A bayesian approach to adaptive video super resolution. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 209–216. IEEE, 2011.

[LS14] Ce Liu and Deqing Sun. On bayesian adaptive video super resolution. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):346–360, 2014.

[LSF07] Ce Liu, Heung-Yeung Shum, and William T Freeman. Face hallucination: Theory and practice. *International Journal of Computer Vision*, 75(1):115–134, 2007.

[LSZ01] Ce Liu, Heung-Yeung Shum, and Chang-Shui Zhang. A two-step approach to hallucinating faces: global parametric model and local nonparametric model. In *Computer Vision and Pattern Recognition*, volume 1, pages I–192. IEEE, 2001.

[LTH+16] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016.

[MC10] Reza Farrahi Moghaddam and Mohamed Cheriet. A multi-scale framework for adaptive binarization of degraded document images. *Pattern Recognition*, 43(6):2186–2198, 2010.

[MI13] Tomer Michaeli and Michal Irani. Nonparametric blind super-resolution. In *IEEE International Conference on Computer Vision*, pages 945–952, 2013.

[Mir05] Boris Mirkin. *Clustering for data mining*. CRC Press, 2005.

[MS01] Sanjit Kumar Mitra and Giovanni L Sicuranza. *Nonlinear image processing*. Academic Press, 2001.

[MTM05] Céline Mancas-Thillou and Majid Mirmehdi. Super-resolution text using the teager filter. In *First International Workshop on Camera-Based Document Analysis and Recognition*. Citeseer, 2005.

[MZQ10] Xiang Ma, Junping Zhang, and Chun Qi. Hallucinating face by position-patch. *Pattern Recognition*, 43(6):2224–2236, 2010.

[NCGKO14] Nibal Nayef, Joseph Chazalon, Petra Gomez-Krämer, and Jean-Marc Ogier. Efficient example-based super-resolution of single text images based on selective patch processing. In *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*, pages 227–231. IEEE, 2014.

[NM14] Kamal Nasrollahi and Thomas B Moeslund. Super-resolution: a comprehensive survey. *Machine vision and applications*, 25(6):1423–1468, 2014.

[NMG+10] Juhan Nam, Gautham J Mysore, Joachim Ganseman, Kyogu Lee, and Jonathan S Abel. A super-resolution spectrogram using coupled plca. In

Eleventh Annual Conference of the International Speech Communication Association, 2010.

[NS08] M.R. Naphade and J.R. Smith. Method and apparatus for semantic super-resolution of audio-visual data, July 3 2008. US Patent App. 11/619,342.

[NTA13] Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki. High-Frequency Restoration Using Deep Belief Nets for Super-resolution. *International Conference on Signal-Image Technology & Internet-Based Systems*, pages 38–42, dec 2013.

[OSS14] Christian Osendorfer, Hubert Soyer, and Patrick Van Der Smagt. Image Super-Resolution with Fast Approximate Convolutional Sparse Coding. *International Conference on Neural Information Processing*, 2014.

[PCRZ06] Lyndsey C Pickup, David P Capel, Stephen J Roberts, and Andrew Zisserman. Bayesian image super-resolution, continued. In *Advances in Neural Information Processing Systems*, pages 1089–1096, 2006.

[PCRZ09] Lyndsey C Pickup, David P Capel, Stephen J Roberts, and Andrew Zisserman. Bayesian methods for image super-resolution. *The Computer Journal*, 52(1):101–113, 2009.

[PE14] Tomer Peleg and Michael Elad. A statistical prediction model based on sparse representations for single image super-resolution. *IEEE Transactions on Image Processing*, 23(6):2569–2582, 2014.

[PETM09] Matan Protter, Michael Elad, Hiroyuki Takeda, and Peyman Milanfar. Generalizing the nonlocal-means to super-resolution reconstruction. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 18(1):36–51, jan 2009.

[PL08] Jeong-Seon Park and Seong-Whan Lee. An example-based face hallucination method for single-frame, low-resolution facial images. *IEEE Transactions on Image Processing*, 17(10):1806–1816, 2008.

[Pla99] N Plaziac. Image interpolation using neural networks. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 8(11):1647–51, jan 1999.

[PRZ06] Lyndsey C Pickup, Stephen J Roberts, and Andrew Zisserman. Optimizing and learning for super-resolution. In *British Machine Vision Conference*, 2006.

[RDS+15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[RN96] Stephen V Rice and Thomas A Nartker. The isri analytic tools for ocr evaluation, 1996.

[RSRB15] Gernot Riegler, Samuel Schulter, Matthias Ruther, and Horst Bischof. Conditioned regression models for non-blind single image super-resolution. In *IEEE International Conference on Computer Vision*, pages 522–530, 2015.

[RZE08] Ron Rubinstein, Michael Zibulevsky, and Michael Elad. Efficient implementation of the k-svd algorithm using batch orthogonal matching pursuit. Technical report, CS - Technion Israel Institute of Technology, 2008.

[SA14] Abhishek Singh and Narendra Ahuja. Super-resolution using sub-band self-similarity. In *Asian Conference on Computer Vision*, pages 552–568. Springer, 2014.

[SCXP15] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 769–777, 2015.

[SE15] Wen-Ze Shao and Michael Elad. Simple, accurate, and robust nonparametric blind super-resolution. In *International Conference on Image and Graphics*, pages 333–348. Springer, 2015.

[SG07] Zohra Saidane and Christophe Garcia. Automatic scene text recognition using a convolutional neural network. In *International Workshop on Camera-Based Document Analysis and Recognition*. Citeseer, 2007.

[SH12] Libin Sun and James Hays. Super-resolution from internet-scale scene matching. In *Computational Photography (ICCP), 2012 IEEE International Conference on*, pages 1–12. IEEE, 2012.

[Sha49] Claude Elwood Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949.

[SKHS98] Toshio Sato, Takeo Kanade, Ellen K Hughes, and Michael A Smith. Video ocr for digital news archive. In *Content-Based Access of Image and*

Video Database, 1998. Proceedings., 1998 IEEE International Workshop on, pages 52–60. IEEE, 1998.

[Sun08] Jian Sun. Image super-resolution using gradient profile prior. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, jun 2008.

[SXS11] Jian Sun, Zongben Xu, and Heung-yeung HY Shum. Gradient profile prior and its applications in image super-resolution and enhancement. *IEEE Transactions on Image Processing*, 20(6):1529–1542, 2011.

[SZ14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[SZT10] Jian Sun, Jiejie Zhu, and Marshall F Tappen. Context-constrained hallucination for image super-resolution. In *Computer Vision and Pattern Recognition (CVPR)*, pages 231–238. IEEE, 2010.

[TB06] Michael E Tipping and Christopher M Bishop. Bayesian image super resolution, September 12 2006. US Patent 7,106,914.

[TC00] Paul D Thouin and Chein-I Chang. A method for restoration of low-resolution document images. *International Journal on Document Analysis and Recognition*, 2(4):200–210, 2000.

[TDG13] Radu Timofte, Vincent De, and Luc Van Gool. Anchored Neighborhood Regression for Fast Example-Based Super-Resolution. *IEEE International Conference on Computer Vision*, pages 1920–1927, dec 2013.

[TDSVG14] Radu Timofte, Vincent De Smet, and Luc Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *Asian Conference on Computer Vision*, pages 111–126. Springer, 2014.

[TLBL10] Yu-Wing Tai, Shuaicheng Liu, Michael S. Brown, and Stephen Lin. Super resolution using edge prior and single image detail synthesis. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2400–2407, jun 2010.

[Tur90] Ken Turkowski. Filters for common resampling tasks. In *Graphics gems*, pages 147–165. Academic Press Professional, Inc., 1990.

[WDA+14] Rim Walha, Fadoua Drira, Adel M Alimi, Frank Lebourgeois, and Christophe Garcia. A sparse coding based approach for the resolution enhancement and restoration of printed and handwritten textual images. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 696–701. IEEE, 2014.

[WDL+13] Rim Walha, Fadoua Drira, Franck Lebourgeois, Christophe Garcia, and Adel M Alimi. Multiple learned dictionaries based clustered sparse coding for the super-resolution of single text image. In *International Conference on Document Analysis and Recognition*, pages 484–488. IEEE, 2013.

[WDL+15] Rim Walha, Fadoua Drira, Frank Lebourgeois, Christophe Garcia, and Adel M Alimi. Resolution enhancement of textual images via multiple coupled dictionaries and adaptive sparse representation selection. *International Journal on Document Analysis and Recognition (IJDAR)*, 18(1):87–107, 2015.

[WDL+16] Rim Walha, Fadoua Drira, Frank Lebourgeois, Adel M Alimi, and Christophe Garcia. Resolution enhancement of textual images: a survey of single image-based methods. *IET Image Processing*, 10(4):325–337, 2016.

[WLY+15] Zhaowen Wang, Ding Liu, Jianchao Yang, Wei Han, and Thomas Huang. Deep networks for image super-resolution with sparse prior. In *IEEE International Conference on Computer Vision*, pages 370–378, 2015.

[WYW+15] Zhangyang Wang, Yingzhen Yang, Zhaowen Wang, Shiyu Chang, Wei Han, Jianchao Yang, and Thomas Huang. Self-tuned deep super resolution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2015.

[WZLP12] Shenlong Wang, Lei Zhang, Yan Liang, and Quan Pan. Semi-Coupled Dictionary Learning with Applications to Image Super-Resolution and Photo-Sketch Synthesis. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2216–2223, 2012.

[YB08] Jiangang Yu and Bir Bhanu. Super-resolution of facial images in video with expression changes. In *Advanced Video and Signal Based Surveillance, 2008. AVSS'08. IEEE Fifth International Conference on*, pages 184–191. IEEE, 2008.

[YFY+16] Wenhan Yang, Jiashi Feng, Jianchao Yang, Fang Zhao, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep edge guided recurrent residual learning for image super-resolution. *arXiv preprint arXiv:1604.08671*, 2016.

[YHY10] Chih-Yuan Yang, Jia-Bin Huang, and Ming-Hsuan Yang. Exploiting self-similarities for single frame super-resolution. In *Asian Conference on Computer Vision*, pages 497–510. Springer, 2010.

[YMY14] Chih-Yuan Yang, Chao Ma, and Ming-Hsuan Yang. Single-image super-resolution: A benchmark. In *European Conference on Computer Vision*, pages 372–386. Springer, 2014.

[YTMH08] Jianchao Yang, Hao Tang, Yi Ma, and Thomas Huang. Face hallucination via sparse coding. In *IEEE International Conference on Image Processing*, pages 1264–1267. IEEE, 2008.

[YWHM10] Jianchao Yang, John Wright, Thomas T.S. Huang, and Yi Ma. Image Super-Resolution via Sparse Representation. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 19(11):1–13, nov 2010.

[YWL+12] Jianchao Yang, Zhaowen Wang, Zhe Lin, Scott Cohen, and Thomas Huang. Coupled dictionary training for image super-resolution. *IEEE transactions on image processing*, 21(8):3467–3478, 2012.

[YWMH08] Jianchao Yang, John Wright, Yi Ma, and Thomas Huang. Image super-resolution as sparse representation of raw image patches. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, jun 2008.

[YY13] Chih-Yuan Yang and Ming-Hsuan Yang. Fast Direct Super-Resolution by Simple Functions. *IEEE International Conference on Computer Vision*, pages 561–568, dec 2013.

[ZEP10] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse representation. In *European Conference on Computer Vision*, 2010.

[ZFC+15] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Learning face hallucination in the wild. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[ZLLT16] Shizhan Zhu, Sifei Liu, Chen Change Loy, and Xiaoou Tang. Deep cascaded bi-network for face hallucination. In *European Conference on Computer Vision*, pages 614–630. Springer, 2016.

[ZP02] Liming Zhang and Fengzhi Pan. A new method of images super-resolution restoration by neural networks. In *Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on*, volume 5, pages 2414–2418. IEEE, 2002.

[ZWTZ16] Xiaole Zhao, Yadong Wu, Jinsha Tian, and Hongying Zhang. Single image super-resolution via blind blurring estimation and anchored space mapping. *Computational Visual Media*, 2(1):71–85, 2016.

FOLIO ADMINISTRATIF

THESE DE L'UNIVERSITE DE LYON OPEREE AU SEIN DE L'INSA LYON

NOM : PEYRARD DATE de SOUTENANCE : 29/09/2017

Prénoms : Clément André

TITRE : Single Image Super-Resolution based on Neural Networks for text and face recognition

NATURE : Doctorat Numéro d'ordre : 2017LYSEI083

Ecole doctorale : InfoMaths (ED 512)

Spécialité : Informatique

RESUME :

Cette thèse porte sur les méthodes de Super-Résolution (SR) pour l'amélioration des performances des systèmes de reconnaissance automatique (OCR, reconnaissance faciale).
Les méthodes de Super-Résolution permettent de générer des images haute résolution (HR) à partir d'images basse résolution (BR). Contrairement à un rééchantillonage par interpolation, elles restituent les hautes fréquences spatiales et compensent les artéfacts (flou, crénelures). Parmi elles, les méthodes d'apprentissage automatique telles que les réseaux de neurones artificiels permettent d'apprendre et de modéliser la relation entre les images BR et HR à partir d'exemples.
Ce travail démontre l'intérêt des méthodes de SR à base de réseaux de neurones pour les systèmes de reconnaissance automatique. Les réseaux de neurones à convolutions sont particulièrement adaptés puisqu'ils peuvent être entraînés à extraire des caractéristiques non-linéaires bidimensionnelles pertinentes tout en apprenant la correspondance entre les espaces BR et HR.
Sur des images de type documents, la méthode proposée permet d'améliorer la précision en reconnaissance de caractère de +7.85 points par rapport à une simple interpolation. La création d'une base d'images annotée et l'organisation d'une compétition internationale (ICDAR2015) ont souligné l'intérêt et la pertinence de telles approches. Pour les images de visages, les caractéristiques faciales sont cruciales pour la reconnaissance automatique. Une méthode en deux étapes est proposée dans laquelle la qualité de l'image est d'abord globalement améliorée, pour ensuite se focaliser sur les caractéristiques essentielles grâce à des modèles spécifiques. Les performances d'un système de vérification faciale se trouvent améliorées de +6.91 à +8.15 points.
Enfin, pour le traitement d'images BR en conditions réelles, l'utilisation de réseaux de neurones profonds permet d'absorber la variabilité des noyaux de flous caractérisant l'image BR, et produire des images HR ayant des statistiques naturelles sans connaissance du modèle d'observation exact.

MOTS-CLÉS : Super-Resolution, Machine Learning, Artificial Neural Networks,

Laboratoire de recherche : LIRIS (UMR 5205)

Directeur de thèse: GARCIA Christophe

Président de jury : M. VIARD-GAUDIN, Christian

Composition du jury :

M. CHATEAU, Thierry	PRU, Univ. de Clermont-Auvergne	Rapporteur
M. THIRAN, Jean-Philippe	PRU, EPFL	Rapporteur
MME GUILLEMOT, Christine	Directrice de recherche, INRIA	Examinatrice
M. VIARD-GAUDIN, Christian	PRU, Univ. de Nantes	Examinateur
M. GARCIA, Christophe	PRU, INSA de Lyon	Directeur de thèse
M. BACCOUCHE, Moez	Dr, Ingénieur de recherche, Orange	Co-encadrant
M. MAMALET, Franck	Dr, Responsable R&D, Spikenet Technologies	Invité