SYSTEMATIC EVALUATION OF SUPER-RESOLUTION USING CLASSIFICATION

Vinay P. Namboodiri[1], Vincent De Smet[1] and Luc Van Gool[1,2]
[1]ESAT-PSI/IBBT, K.U.Leuven, Belgium
[2] Computer Vision Laboratory, BIWI/ETH Zürich, Switzerland

Abstract— **Currently two evaluation methods of super-resolution (SR) techniques prevail: The objective Peak Signal to Noise Ratio (PSNR) and a qualitative measure based on manual visual inspection. Both of these methods are sub-optimal: The latter does not scale well to large numbers of images, while the former does not necessarily reflect the perceived visual quality. We address these issues in this paper and propose an evaluation method based on image classification. We show that perceptual image quality measures like structural similarity are not suitable for evaluation of SR methods. On the other hand a systematic evaluation using large datasets of thousands of real-world images provides a consistent comparison of SR algorithms that corresponds to perceived visual quality. We verify the success of our approach by presenting an evaluation of three recent super-resolution algorithms on standard image classification datasets.**

I. Introduction

Super-resolution (SR) of images and video is an area of active research that has been extensively studied [1], [2], [3], [4], [5], [6]. Various approaches have been proposed: interpolation, reconstruction and learning-based methods. Among these, the learning-based methods [2] are currently the ones achieving highest qualities of super-resolution [3].

However, despite the wide variety of different super-resolution methods, there has been significantly less investigation into methods for their evaluation and comparison. One often used evaluation criterion is the manual inspection and subjective evaluation of the visual quality of the super-resolved images [7], [8]. This approach is labour intensive and thus restricted to limited numbers of result images. Other popular evaluation measures are the Peak Signal to Noise Ratio (PSNR) and the Mean Squared Error (MSE). These are objective measures that scale well to large numbers of images, however, they do not necessarily correspond to perceived visual quality.

Perceptual image quality assessment has received attention and several measures have been proposed, such as Structural Similarity (SSIM) [9], information theoretic [10] and wavelet-based methods [11]. These are however reference based, i.e. they need an undistorted, ideal version of the image as reference. Wang *et al.* [12] proposed an approach for no-reference quality assessment which was tuned towards estimating the blocky artifacts of JPEG compression. Recently there have been two more notable approaches. Moorthy and Bovik [13] explicitly model and attempt to classify the distortion. Another approach uses statistics of the DCT representation of an image to estimate its quality [14]. One common feature of these explicit approaches is that they are more suited to estimate image quality in the presence of severe image distortions. The subtle differences that are present in the outputs of super-resolution algorithms cannot be reliably estimated by these approaches. We propose a more robust setting using an explicit cognition-based approach towards evaluation of SR.

Furthermore, all these evaluation methods do not take into account the differences between super-resolution and general image enhancement, which often aims at denoising and deblurring. Super-resolution on the other hand aims more at synthesizing high-resolution details which might not at all be observable in the low-resolution image. However, there is no inherent guarantee that a super-resolved image retains the same semantic meaning. This is especially a problem for the learning-based methods and the objective image quality measures mentioned above are unlikely to detect such errors. Hence we argue the necessity for evaluation methods on a higher visual level. In this context we propose to utilize image classification techniques [15], [16], which try to classify whole objects or scenes. Lately, substantial progress for object classification in real-world images has been shown [17]. In this paper, we use a state-of-the-art image classification system based on Locality-constrained Linear Coding (LLC) [18].

Our main contribution in this paper is a new systematic and objective measure for the evaluation of super-resolution methods. This measure does not require the ground-truth as reference. Our evaluation suggests that the proposed method is consistent in its ranking of super-resolution algorithms with visual perception. We argue that using object classification techniques based on local features closely resembles our visual perception of image quality. The proposed method does not only provide a relative ordering of super-resolution techniques, but also an idea of their absolute performance. Furthermore, the classification results provide us with a more objective measure than the manual visual inspection of example images. In particular, it allows us to evaluate super-resolution techniques on large collections of images, like the ones used for object detection [17], [19], [20]. We will provide results in Section IV that demonstrate the success of our approach.

II. Learning-based super-resolution

A. Related Work

Learning-based super-resolution methods almost always employ a database or dictionary created from corresponding

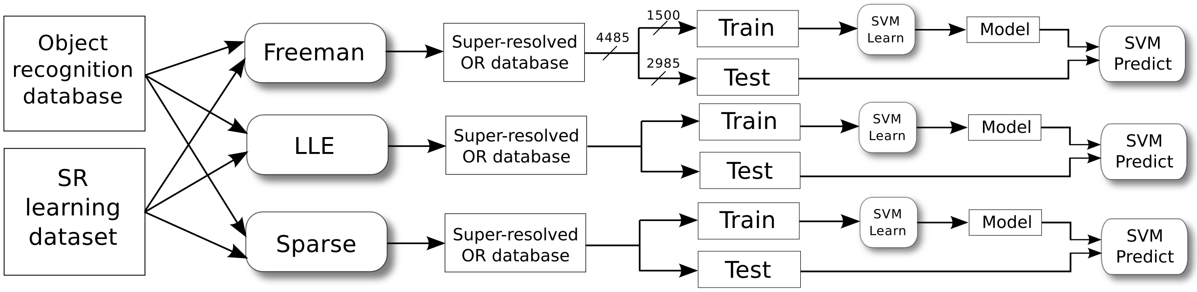

Fig. 1. Illustration of the processing pipeline for evaluation of super-resolution.

low resolution (LR) and high resolution (HR) image pairs. One of the foremost papers written on this approach is the patch-based work by Freeman *et al.* [2]. The authors propose the use of image patch exemplars for learning the relationship between low- and high-resolution images in a Markov Random Field (MRF) framework. This idea has been extended by the authors in [21] with an approximate one-pass solution to the Markov network. A similar approach has been proposed by Baker and Kanade [4], where low level features are recognized and the corresponding high resolution features are hallucinated. Their method is geared towards specific scene content, such as faces and text. There have been further extensions to the Freeman method, for instance by Wang *et al.* [22] where the authors add mutual co-occurrence information to the model. Sun *et al.* propose a gradient prior [23], learned from edge statistics in natural images, to enforce a gradient field constraint. Adding context to learning-based super-resolution has also been studied recently by Sun *et al.* [24] and by HaCohen *et al.* [25]. While the patch-based methods discussed previously use the entire database of LR-HR image pairs, there have been dictionary-based methods that learn a compact dictionary from the database [5], [6]. The advantage of these dictionary learning approaches is that they scale well with increased sizes of the database.

B. Evaluated super-resolution algorithms

We have implemented the learning-based methods by Freeman *et al.* [2] and the Locally Linear Embedding (LLE) method by Chang *et al.* [5]. We compare these methods with the recent work on sparse representation for super-resolution by Yang *et al.* [6]. In the Freeman method, band-pass filtered patches are obtained from the LR-HR database. In the original work, the authors use a B-tree representation to efficiently search for image patches. We have implemented even more efficient data-structures based on adaptive Locality Sensitive Hashing (LSH) [22] to retrieve the patches. We have also experimentally compared MRF-based belief propagation techniques with a simplified approach of dense patch sampling and normalization. We have observed that the latter approach gives equivalent results to the MRF-based approach and we therefore adopt this approach.

Patch-based methods depend on the size of the learning database. The approach proposed by Chang *et al.* [5] does not depend on this size. In their paper the authors use the LLE dimensionality reduction approach to obtain a mapping between LR and HR patches. For each patch of the LR image a set of K nearest neighbors is found in the dictionary. K reconstruction weights are then computed that minimize the reconstruction error for each LR patch. These weights can be used to create a high-resolution patch, effectively approximating the scene patch as a linear combination of database patches, which can be described as points in the lower-dimensional database space.

This approach has been extended by Yang *et al.* [6] by using a sparsity constraint to learn a dictionary. They start from the assumption that an efficient sparse dictionary can be created from patch-based features, randomly sampled from natural images. The HR patches forming the super-resolved image are then created as a sparse linear combination of dictionary patches.

In the next sections, we show how object classification can be used to compare different super-resolution methods. To show this we compare the method proposed by Freeman *et al.* [2], the LLE-based method of Chang *et al.* [5] and the sparse approach of Yang *et al.* [6]. These methods are also compared with the results of bicubic interpolation as a reference.

III. Classification for evaluation

In order to perform classification we use a state-of-the-art object classifier based on Locality-constrained Linear Coding (LLC) [18]. This classifier improves the bag of visual words based spatial pyramid matching method [19] for object classification. The classifier enforces explicit sparsity based on a locality constraint that allows discriminative and computationally efficient classification. The LLC classifier has been shown to perform well on all challenging object classification datasets. The pipeline for classification involves a) computing of dense SIFT features [26] at a fixed scale for all images, b) computing a visual vocabulary from the features of the training images, c) obtaining a locality-constrained linear coding for the SIFT features by projecting them onto the visual vocabulary, d) obtaining a multi-scale spatial max pooling for the features of an image and e) classifying them with support vector machines. While in the original paper the authors use a linear kernel, in our implementation, we use a histogram intersection kernel [19] that shows better accuracy.

Method	NR-JPG 15-scene	Classification 15-scene	NR-JPG TU Graz	Classification TU Graz
Original	9.0463	82.28%	10.6007	88.39%
Bicubic	9.3740	82.68%	11.1791	86.84%
LLE	8.0843	82.78%	10.8031	86.99%
Sparse	7.9211	83.25%	11.0221	86.69%
Freeman	10.3684	83.58%	11.5159	87.62%

TABLE I
CLASSIFICATION ACCURACY AND NR-JPG SCORE FOR UPSAMPLING WITH MAGNIFICATION FACTOR $3\times$.

The procedure for the proposed evaluation of super-resolution using classification is illustrated in Fig. 1. The approach entails learning-based super-resolution of image classification datasets followed by classification. The database used for learning-based super-resolution is the one used by Yang *et al.* [6] which is a general collection of natural images. None of these images are included in the image classification datasets. This enables us to study a general model for super-resolution rather than being tuned towards specific applications like face recognition or text analysis. The evaluation is done on standard image classification datasets by using the procedure described above. The super-resolved images are split into training and testing subsets. For each of these sets the same support vector machine parameters are used. The evaluation measure used is the average classification accuracy per class for the multi-class classification of the test dataset. Higher classification accuracy is obtained when the local features semantically capture the class content reliably. This corresponds to the required goal of evaluation of super-resolution based on visual perception.

IV. EXPERIMENTS

We test the different super-resolution methods on two popular image datasets. The first is the 15-scene category database by Lazebnik *et al.* [19]. The dataset consists of fifteen widely varying scene categories, including bedroom scenes, industrial areas, city scenes and mountains, containing a total of 4485 images. The second database is the TU Graz image database compiled by Opelt and Pinz [20], containing three classes: bikes, cars and people. This dataset contains a total of 1096 images.

In Table I, we present the results for both datasets. We compare our results to the no-reference JPEG (NR-JPG) quality measure [12]. All images are upsampled by a factor of 3 using each of the three super-resolution algorithms. We additionally provide results for the original low-resolution image and bicubic interpolation. We use these results as an input for the classifier. In the case of the 15-scene category dataset, we train on 100 images for each scene category and test on the rest, resulting in 1500 training images and 2985 test images. We use the same approach for the TU Graz dataset. The average classification accuracy for these images is shown in Table I. The classification evaluation results are always consistent with perceived visual quality and in the ranking of super-resolution algorithms, while the NR-JPG method is not. An interesting observation is that the NR-JPG method perceives bicubic interpolation as being better than the super-resolution algorithms. However, visually the super-resolution algorithms clearly perform better. All tested super-resolution algorithms give increased accuracy in classification with respect to the original images and bicubic interpolation for the 15-scene category dataset. For the TU Graz dataset, the classification does not improve over the original, but does improve over bicubic interpolation. This can be explained by considering that we use a fixed scale for the local features for the two datasets. However, the scale of the underlying scene features for TU Graz is already large, so the $3\times$ zoom makes the extracted features less useful for the classifier. The modification of the scale of features is interesting but not considered in the scope of the present work. Among the various algorithms, the method by Freeman *et al.* gives the best improvement in classification accuracy. This observation can be verified through visual inspection of the resulting images. We have included some examples in Fig. 2. These are cropped parts of images from both datasets. Additional results of each super-resolution algorithm for all dataset classes can be found on our website[1].

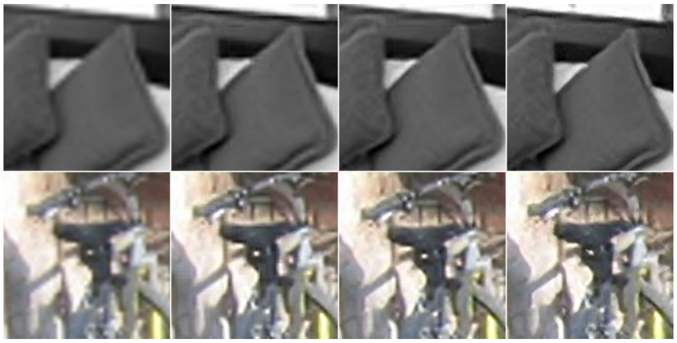

Fig. 2. Some cropped examples from the 15-scene dataset (top) and the TU Graz dataset (bottom). The columns show different methods ordered by perceived visual quality (left-right): bicubic, LLE, sparse, Freeman.

An alternative way in which classification can be used is by downsampling the original image and upsampling it with the various super-resolution algorithms. This allows a comparison between the standard evaluation methods of PSNR, SSIM and our proposed method of using the classification accuracy. As the sparse representation method requires a computationally very expensive retraining of the dictionary, this comparison has been done only for the method of Freeman, the LLE method and the bicubic algorithm. The evaluation allows us to compare the visual accuracy with both the reference and no-reference methods and the classification accuracy. The results are shown in Table II. The comparison demonstrates the drawbacks of common image quality measures for evaluation of super-resolution. These measures are inconsistent in the ranking of super-resolution algorithms. For instance, PSNR considers the Freeman method to be better than bicubic interpolation for the 15-scene dataset, however, it considers vice-versa for the TU

[1] http://homes.esat.kuleuven.be/∼vdesmet/sr_eval/

Method	PSNR 15-scene	SSIM 15-scene	NR-JPG 15-scene	Classification 15-scene	PSNR TU Graz	SSIM TU Graz	NR-JPG TU Graz	Classification TU Graz
Original	-	-	9.0463	82.28%	-	-	10.6007	88.39%
Downsampled	-	-	9.5529	76.52%	-	-	10.6459	86.99%
Bicubic	26.36	0.8701	9.7522	81.04%	29.75	0.9491	10.8327	86.99%
LLE	25.92	0.8535	9.4349	81.57%	28.33	0.9404	10.4041	87.62%
Freeman	27.17	0.8560	9.9730	82.24%	28.36	0.9394	10.4656	87.62%

TABLE II

A comparison of image quality measures with the proposed method when the images are first downsampled before being super-resolved to their original size. The magnification factor is $2\times$.

Graz dataset. Similarly, SSIM considers the Freeman method to be better than LLE for the 15-scene dataset, but the opposite for the TU-Graz dataset. The ranking of the super-resolution algorithms are also not consistent with visual perception. This was reported by Reibman *et al.* [7] as well, where they opted for subjective measures. The proposed classification method provides a consistent, perceptual objective measure for comparison of super-resolution algorithms. An additional observation is that the classification measure of the Freeman algorithm differs only by 0.04% from the original images. This provides an objective measure of perceptual proximity of the super-resolved results to the original high-resolution images.

V. Conclusion

In this paper we propose a new way to systematically evaluate super-resolution algorithms. We achieve this by first super-resolving standard image classification datasets and then classifying them with a state-of-the-art classification algorithm. The use of distinctive local features during classification ensures closer correspondence with perceptual visual quality. This provides us with a consistent objective measure better suited for evaluating super-resolution than widely used image quality measures like PSNR and SSIM and no-reference techniques like NR-JPG. The proposed evaluation scheme works on a much larger scale than manual visual inspection. The resulting image quality evaluation is therefore an apt measure of the performance of a super-resolution algorithm.

References

[1] Peyman Milanfar, *Super-Resolution Imaging*, Taylor & Francis/CRC Press. Series: Digital Imaging and Computer Vision, 2010.

[2] William T. Freeman, Egon C. Pasztor, and Owen T. Carmichael, “Learning low-level vision,” *International Journal of Computer Vision*, vol. 40, no. 1, pp. 25–47, 2000.

[3] Zhouchen Lin, Junfeng He, Xiaoou Tang, and Chi-Keung Tang, “Limits of learning-based superresolution algorithms,” *International Journal of Computer Vision*, vol. 80, pp. 406–420, 2008.

[4] Simon Baker and Takeo Kanade, “Hallucinating faces,” in *FG*. 2000, pp. 83–89, IEEE Computer Society.

[5] Hong Chang, Dit-Yan Yeung, and Yimin Xiong, “Super-resolution through neighbor embedding,” in *CVPR*, 2004, vol. 1, pp. 275 –282.

[6] Jianchao Yang, J. Wright, T.S. Huang, and Yi Ma, “Image super-resolution via sparse representation,” *Image Processing, IEEE Transactions on*, vol. 19, no. 11, pp. 2861 –2873, 2010.

[7] Amy R. Reibman, Robert M. Bell, and Sharon Gray, “Quality assessment for super-resolution image enhancement,” in *ICIP*, 2006, pp. 2017–2020.

[8] Amy R. Reibman and Thilo Schaper, “Subjective performance evaluation of super resolution image enhancement,” in *Proceedings of the Second International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2006.

[9] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[10] Hamid R. Sheikh, Alan C. Bovik, and Gustavo de Veciana, “An information fidelity criterion for image quality assessment using natural scene statistics,” *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2117–2128, 2005.

[11] Zhou Wang and Alan C. Bovik, “Embedded foveation image coding,” *IEEE Transactions on Image Processing*, vol. 10, no. 10, pp. 1397–1410, 2001.

[12] Zhou Wang, Hamid R. Sheikh, and Alan C. Bovik, “No-reference perceptual quality assessment of jpeg compressed images,” in *ICIP (1)*, 2002, pp. 477–480.

[13] A. K. Moorthy and A. C. Bovik, “A two-step framework for constructing blind image quality indices,” *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 587–599, 2010.

[14] Michele A. Saad, Alan C. Bovik, and Christophe Charrier, “Natural dct statistics approach to no-reference image quality assessment,” in *ICIP*, 2010, pp. 313–316.

[15] Axel Pinz, “Object categorization,” *Foundations and Trends in Computer Graphics and Vision*, vol. 1, no. 4, 2005.

[16] Jean Ponce, Martial Hebert, Cordelia Schmid, and Andrew Zisserman, Eds., *Toward Category-Level Object Recognition*, vol. 4170 of *Lecture Notes in Computer Science*. Springer, 2006.

[17] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results,” .

[18] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas S. Huang, and Yihong Gong, “Locality-constrained linear coding for image classification,” in *CVPR*, 2010, pp. 3360–3367.

[19] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *CVPR*, 2006, vol. 2, pp. 2169 – 2178.

[20] Andreas Opelt, Axel Pinz, Michael Fussenegger, and Peter Auer, “Generic object recognition with boosting,” *PAMI*, vol. 28, pp. 2006, 2004.

[21] William T. Freeman, Thouis R. Jones, and Egon C. Pasztor, “Example-based super-resolution,” *IEEE Computer Graphics and Applications*, vol. 22, no. 2, pp. 56–65, 2002.

[22] Qiang Wang, Xiaoou Tang, and Harry Shum, “Patch based blind image super resolution,” in *ICCV*, 2005, pp. 709–716.

[23] Jian Sun, Zongben Xu, and Heung-Yeung Shum, “Image super-resolution using gradient profile prior,” in *CVPR*, 2008.

[24] Jian Sun, Jiejie Zhu, and Marshall F. Tappen, “Context-constrained hallucination for image super-resolution,” in *CVPR*, 2010, pp. 231–238.

[25] Y. HaCohen, R. Fattal, and D. Lischinski, “Image upsampling via texture hallucination,” in *Computational Photography (ICCP)*, march 2010, pp. 1 –8.

[26] David G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.