QUANTITATIVE ANALYSIS OF RESOLUTION SYNTHESIS

Ramez Yoakeim and David Taubman

The University of New South Wales, Sydney, Australia

ABSTRACT

We address a number of gaps left by recent work on resolution synthesis image interpolation, including the validity of some assumptions which we examine and verify. The relative merit of a mixture approach versus a maximum likelihood approach, the impact of the classification process, and the impact of the number of classes on the performance of the interpolator are all investigated. We also examine the suitability of the underlying statistical models. We propose a modified synthesis component, based on the discrete wavelet transform, as an alternative to the non-overlapped block synthesis process described in previous work. We also introduce a number of measures to significantly improve the computational efficiency and suitability of RS for automated, unassisted classification training.

1. INTRODUCTION

Resolution Synthesis (RS) describes a class of algorithms that aim to estimate a higher resolution image, given a single low resolution instance of the same image. Resolution Synthesis fits within the broader class of problems known as "Inverse Problems in Imaging," which includes related ill-posed inverse problems such as super-resolution, deblurring, denoising and other restoration objectives. While various approaches to resolution synthesis have been proposed [1, 2], in this paper we focus on the approach introduced by Atkins et al. [3, 4].

This approach consists of two key elements: a classification strategy for the low resolution image pixels; and a high resolution synthesis (interpolation) strategy, which is based on the classification. Specifically, the neighborhood (typically 5×5 pixels) surrounding each low resolution image pixel is interpreted as a realization of one of M Gaussian generators, corresponding to M hidden class models which represent oriented edges, texture and the like. The particular class, j, to which a neighborhood belongs cannot generally be determined unambiguously, based solely on the observed pixel values. However, the observed pixel values can be used to deduce a posterior class membership probability distribution, based on the individual class probability distributions, and a set of prior class likelihoods. To produce the high resolution image, Atkins et al. employ a collection of LMMSE (Linear Minimum Mean Squared Error) estimators, one for each class, taking the expectation of these estimator outputs over the posterior class distribution.

We provide further details concerning the Gaussian class models and LMMSE estimators in Section 2. For the moment, however, we note that the determination of their parameters is the subject of training. The EM algorithm is leveraged to find a set of class models which best explain the features of a collection of low resolution training images. Each low resolution training image has an associated, known, high resolution image. Classical estimation theory is then used to derive the LMMSE interpolators, based on a maximum likelihood classification of the neighborhoods found in each low resolution training image, and the cross-correlation between these classified neighborhood pixels and the corresponding high resolution image pixels. Various alternatives [5, 4] to the RS approach in [3] focus on improving the computational efficiency of the interpolation process and providing alternate methods for estimating the class models.

In the present paper, we examine several aspects of the classification-based RS paradigm presented in [3]. It is natural to ask how much of the performance of the algorithm is attributable to the LMMSE interpolation process and how much additional gain is derived from the classification structure. Indeed, since the classification and synthesis (interpolation) aspects of the algorithm are not tied together within the training process, it is unclear a priori whether classification should improve or degrade the synthesized images. Related questions surround the best number of classes to use. We also investigate the question of whether the image is best synthesized by taking the expectation of the LMMSE estimates over the posterior class distribution at each location, or by using the single LMMSE estimate corresponding to the most likely class (mode of the posterior distribution) at each location. Reassuringly, we find that the former approach (the one proposed by Atkins et al.) is indeed superior. By and large, these matters do not appear to have been previously addressed.

In addition to a more careful investigation into important aspects of the existing algorithm, we propose several novel variations here. We observe that the originally proposed Gaussian class models are not particularly effective in discriminating between semantically recognizable image features. Based on this observation, we propose the use of Gaussian class models which can fully exploit the covariance structure found within each low resolution pixel neighborhood. We report on a novel population subsampling strategy to dramatically improve the efficiency of the EM training procedure. Finally, we propose a modified synthesis (interpolation) stage, utilising the Discrete Wavelet Transform (DWT) to effect overlapping, rather than blocked synthesis kernels.

The paper is organised as follows. Section 2 provides an overview of the classification based RS algorithm, including a discussion of its underlying hypotheses. Section 3 discusses population subsampling to reduce the computational burden of training. Section 4 provides experimental evidence to evaluate the various hypotheses mentioned in Section 2. Then Section 5 presents an alternate interpolation strategy within the same framework of resolution synthesis.

2. CLASSIFICATION BASED RESOLUTION SYNTHESIS

In this section, we provide further details concerning the classification and synthesis elements of the RS algorithm. We write

0-7803-8554-3/04/$20.00 ©2004 IEEE.

$x[\mathbf{n}] \equiv x[n_1, n_2]$ for the high resolution image which gives rise to a corresponding low resolution image $u[\mathbf{n}]$. The framework makes no specific assumptions concerning the resolution reduction process. However, in order to train the model parameters, low resolution resolution images must be generated from an initial set of high resolution training images. For this purpose we adopt a conventional filtering and subsampling process, with a sampling ratio of 2 and

$$u[\mathbf{n}] = \sum_{\mathbf{k}} h[\mathbf{k}]\, x[2\mathbf{n} - \mathbf{k}]$$

In the simple case of 2×2 averaging,

$$h[\mathbf{k}] = \begin{cases} 1 & k_1, k_2 \in \{0, -1\} \\ 0 & \text{otherwise} \end{cases}$$

This is the example adopted in [3], although many imaging processes involve significantly smoother blurring kernels.

Associated with each low resolution image pixel $u[\mathbf{n}]$, we may identify a 5×5 neighbourhood $\mathbf{z}[\mathbf{n}]$, consisting of the image samples $u[\mathbf{n} + \mathbf{k}]$, $-2 \leq k_1, k_2 \leq k$, arranged in vector form for convenience. For the purpose of Gaussian statistical modeling, Atkins et al. suggest the following non-linear transformation of $\mathbf{z}[\mathbf{n}]$ to an 8-dimensional vector, $\mathbf{y}[\mathbf{n}]$:

$$\mathbf{y} = \begin{cases} \mathbf{y}' \cdot \|\mathbf{y}'\|^{-3/4} & \mathbf{y}' \neq \mathbf{0} \\ \mathbf{0} & \text{otherwise} \end{cases} \tag{1}$$

Here, $\mathbf{y}'[\mathbf{n}]$ is obtained by subtracting $u[\mathbf{n}]$ from its 8 immediate neighbours.

We follow the usual convention of identifying random variables and vectors by upper case versions of their deterministic counterparts. The mean-removed, non-linearly transformed neighbourhood, $\mathbf{Y}[\mathbf{n}]$, is modeled as a multivariate Gaussian mixture, with PDF

$$p_{\mathbf{Y}}(\mathbf{y}) = \sum_{j=1}^{M} \pi_j p_{\mathbf{Y}|J}(\mathbf{y}, j) \tag{2}$$

Here, j denotes one of M underlying classes, π_j is the prior probability that $J = j$, and

$$p_{\mathbf{Y}|J}(\mathbf{y}, j) = \frac{1}{\left(\sqrt{2\pi}\right)^8} e^{-\frac{1}{2\sigma^2}\|\mathbf{y} - \boldsymbol{\mu}_j\|^2} \tag{3}$$

Note that each element of $\mathbf{Y}[\mathbf{n}]$ is being modeled here as an independent Gaussian random variable, with a unique class-dependent mean, and a common class-independent variance, σ^2.

Given an observed low resolution image $u[\mathbf{n}]$, we compute $\mathbf{y}[\mathbf{n}]$ for each $\mathbf{n}$, and evaluate the posterior likelihood that the neighbourhood $\mathbf{z}[\mathbf{n}]$ was generated by class j, for each $j = 1, 2, \ldots M$. Applying Bayes rule, we find that

$$p_{J|\mathbf{Z}}(j, \mathbf{z}) = p_{J|\mathbf{Y}}(j, \mathbf{y}) = \frac{\pi_j \exp\left(\frac{-1}{2\sigma^2}\left\|\mathbf{y} - \boldsymbol{\mu}_j\right\|^2\right)}{\sum_{l=1}^{M} \pi_l \exp\left(\frac{-1}{2\sigma^2}\left\|\mathbf{y} - \boldsymbol{\mu}_l\right\|^2\right)} \tag{4}$$

where the first equality holds by assumption. The parameters $\boldsymbol{\mu}_j$, π_j and σ are derived by a training process, which aims to maximize

$$\sum_i \log p_{\mathbf{Y}}\left(\mathbf{y}_i | \boldsymbol{\mu}_j, \pi_j, \sigma\right), \tag{5}$$

where the $\mathbf{y}_i$ are drawn from the mean-removed transformed neighbourhoods of a collection of low resolution training images, $u[\mathbf{n}]$. The "expectation maximization" (EM) algorithm is used for this purpose. This explains the classification part of the procedure.

For synthesis (interpolation), we first write $\mathbf{x}[\mathbf{n}]$ for the 2×2 block of samples $x[2\mathbf{n}]$, $x[2n_1 + 1, 2n_2]$, $x[2n_1, 2n_2 + 1]$ and $x[2n_1 + 1, 2n_2 + 1]$, arranged in vector form. This allows us to associate each $\mathbf{x}[\mathbf{n}]$ with a unique low resolution neighbourhood $\mathbf{z}[\mathbf{n}]$. We model the conditional distribution of $\mathbf{X}$ given $\mathbf{Z} = \mathbf{z}$, for each class j, using a multivariate Gaussian, with mean $\mathbf{A}_j\mathbf{z} + \boldsymbol{\beta}_j$. This leads to the following formula for the expected value, $\hat{\mathbf{x}}[\mathbf{n}]$, of $\mathbf{X}[\mathbf{n}]$ given $\mathbf{z}[\mathbf{n}]$

$$\hat{\mathbf{x}}[\mathbf{n}] = \sum_{j=1}^{M} p_{J|\mathbf{Y}}(j, \mathbf{y}[\mathbf{n}]) \cdot \left(\mathbf{A}_j\mathbf{z} + \boldsymbol{\beta}_j\right) \tag{6}$$

Noting that the mode and mean of a Gaussian distribution, we point out that the MAP estimate, $\hat{\mathbf{x}}'[\mathbf{n}]$, is

$$\hat{\mathbf{x}}'[\mathbf{n}] = \mathbf{A}_{j_{\mathbf{n}}}\mathbf{z} + \boldsymbol{\beta}_{j_{\mathbf{n}}}\big|_{j_{\mathbf{n}} = \operatorname{argmax}_j p_{J|\mathbf{Y}}(j, \mathbf{y}[\mathbf{n}])} \tag{7}$$

The parameters, $\mathbf{A}_j$ and $\boldsymbol{\beta}_j$ are derived from a collection of low and high resolution training images, using classical estimation theory.

Examining the brief derivation above, we note the following:

1. The classification objective of equation (5) has no dependence whatsoever on the high resolution image, and hence the synthesis problem itself. While this allows some flexibility in what the classification model is used for, once it has been determined, it does not guarantee optimal utility in the interpolation problem.
2. The choice of a single scalar variance parameter, σ, with independent Gaussians for each element in the mean-removed transformed neighbourhood vector $\mathbf{Y}$, reduces the classification procedure to a pattern matching exercise. The patterns are represented by the class centroids $\boldsymbol{\mu}_j$, while σ controls the degree to which we prefer the best matching pattern over other, more distant matches.
3. The non-linear transformation of equation (1) is not insensitive to variations in luminance for what amount to be the same feature, such as an edge of a specific orientation for example. This results in the creation of redundant classes, which model essentially the same feature with different luminances.

3. TRAINING AND COMPUTATION

As mentioned, model training uses the EM algorithm to maximise (5). This step in the procedure can be very computationally demanding. We note that Atkins et al. [3] recommend the use of $M \approx 100$ classes and 10^5 training samples, $\mathbf{y}_i$. Just the cost of computing distances between each feature vector and class centroid, as per equation (4), at every iteration of the algorithm, makes the computation unwieldy even for experimental purposes.

To reduce the computational requirement, we propose a reduction both in the size of the initial training set and the number of classes. Reducing the size of the initial training set is achieved using a population subsampling technique. We adopt a distance-based clustering approach to pare down an initial set of training vectors, $\mathbf{y}_i$. Specifically, our goal is to replace large numbers of very similar training vectors with a single representative vector. The approach is particularly effective in eliminating redundant

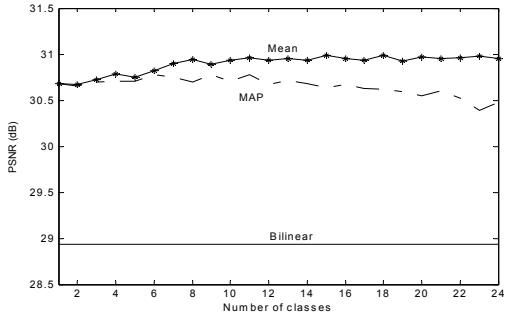

Fig. 1. Results for mean and MAP synthesis

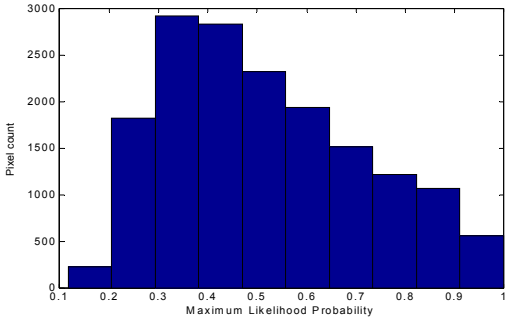

Fig. 2. Probability distribution of MAP estimator

vectors from expansive smooth regions of a training image. To preserve the original probability distribution, each retained vector $\mathbf{y}_i$ is assigned a weight w_i, corresponding to the size of the original cluster which it represents. The weights are then introduced into equation (4) to obtain

$$p_{J|\mathbf{Y}}(j, \mathbf{y}_i) = \frac{w_i \pi_j \exp\left(\frac{-1}{2\sigma^2} \left\|\mathbf{y}_i - \boldsymbol{\mu}_j\right\|^2\right)}{\sum_{l=1}^{M} w_i \pi_l \exp\left(\frac{-1}{2\sigma^2} \left\|\mathbf{y}_i - \boldsymbol{\mu}_l\right\|^2\right)} \qquad (8)$$

Using an iterative split and merge clustering algorithm, a variant of ([6]), we can adjust the size of the subsampled population to establish an arbitrary level of modeling accuracy. We also use the K-means algorithm to create an initial set of classes for the EM algorithm, in preference to the tedious manual initialization suggested in [3]. Our objective is to examine the performance of the interpolation framework in a more realistic setting, where manual intervention in the training set selection and class initialisation processes is unlikely to be possible.

For the purpose of the findings below, we opt to use the same image for both training and testing. Although this is unconventional, we note that it helps us to focus on the inner dynamics of the interpolation approach. Moreover, it allows us to separate the impact of classification and synthesis, since the classification accuracy may be controlled directly by adjusting the size of our subsampled population.

4. QUANTITATIVE ANALYSIS OF RS HYPOTHESES

There are three distinct implicit assumptions underlying the interpolation framework introduced by Atkins et al. [3]. It is reasonable to expect that even with a single class, LMMSE interpolation

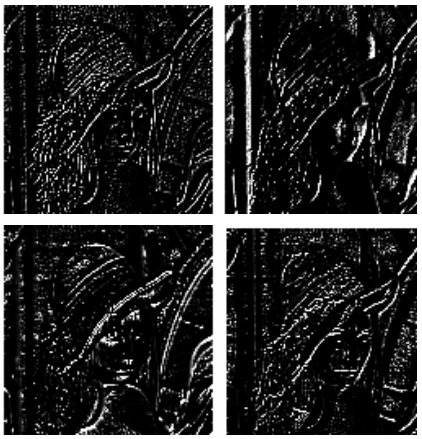

Fig. 3. Four example class maps, out of 24.

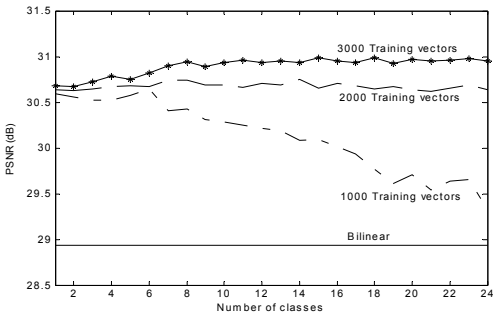

Fig. 4. Results for 3 population sizes

should be superior to classical methods such as bilinear interpolation. The assumption that multiple classes can improve the result, however, needs to be tested. A second assumption is that the mean estimator, $\hat{\mathbf{x}}[\mathbf{n}]$, of equation (6) will produce superior results to the MAP estimator, $\hat{\mathbf{x}}'[\mathbf{n}]$, of equation (7). This assumption also needs to be tested. Finally, we investigate the assumption that the independent Gaussian class model of equation (3) can adequately discriminate between the image features which are important for interpolation.

Figure 1 reveals that the mean estimator $\hat{\mathbf{x}}[\mathbf{n}]$ does indeed outperform the MAP estimator $\hat{\mathbf{x}}'[\mathbf{n}]$. This is reinforced by the fact that, as the number of classes increases, the improvement offered by mean synthesis over MAP synthesis also grows. This is reassuring, especially since the MAP synthesis of equation (7) is a great deal less complex than that of equation (6). It is also reassuring to observe that both estimators outperform classical bilinear interpolation, which is to be expected considering that each class employs a trained LMMSE interpolator.

The difference between mean and MAP synthesis may further be appreciated by considering Figure 2. This figure plots the number of locations $\mathbf{n}$, whose most likely class $j_{\mathbf{n}}$ has probability $p_{J|\mathbf{Y}}(j_{\mathbf{n}}, \mathbf{y}[\mathbf{n}]) \in I_p$, where the intervals I_p serve only to quantize the range of possible maximum likelihoods, shown on the horizontal axis. Evidently, there are many locations for which no one class dominates the mean estimate, $\hat{\mathbf{x}}[\mathbf{n}]$. This behaviour may reflect poorly on the classification process itself.

To investigate the effectiveness of the classification process

more carefully, consider Figure 3. The intensity within each of the four images represents the scaled posterior class membership probability for a particular class. Only four classes are shown here, from a trained set of $M = 24$ classes; however. We can make two observations. The first is that each class supports a wide range of orientations. Our second observation is that individual locations can have a high probably of membership in several classes, even when the classes represent important features such as oriented image edges. Similar conclusions may be drawn by examining other sets of classes from the 24 total considered here.

The relatively poor semantic association revealed by Figure 3 is a direct result of the classification scheme's reliance on class mean patterns $\boldsymbol{\mu}_j$, as the classification differentiator. This builds an intensity dependence into the classes which works against semantic association. In ongoing work, we are examining a more general multivariate Gaussian modeling approach, utilising covariance statistics of the vectors $\mathbf{y}[\mathbf{n}]$ in order to more robustly discriminate between edge features.

We now consider the impact of population subsampling and the number of classes, M, on the effectiveness of the RS interpolator. Figure (4) suggests that little improvement in interpolated image quality is achieved by increasing the number of classes beyond about 10, at least in our experiments. Similar observations hold as the number of classes is allowed to become much larger, approaching the value of 100 recommended in [4]. Interestingly, increasing the number of classes can actually cause the interpolated image quality to degrade. In fact, there is no a priori reason to assume that this should not happen, since the classification parameters are trained without any explicit regard for the synthesis problem.

The impact of population subsampling can be seen more clearly with the selection of relatively small population sizes in Figure 4. As the number of training vectors, $\mathbf{y}_i$, increases, the interpolation quality also improves. Considering the size and representative nature of these training sets, this behaviour suggests once more that the classification strategy is having a hard time discovering semantically meaningful image statistics. Indeed, considering the illumination dependence of the class models, any small deviation between the statistics of the image and those represented by the training vectors can result in the selection of inappropriate MMSE interpolators. These behaviours were consistently observed for a number of input images exhibiting differing mixes of textures and edge orientations. As noted above, these are limitations which we are working to eliminate by modifying the underlying class models.

5. DWT-BASED RESOLUTION SYNTHESIS

One of the structural attributes of the RS approach presented so far is the independent interpolation of each block of 2×2 samples, represented by the vector $\mathbf{x}[\mathbf{n}]$. What we anticipate would provide better outcomes is an approach that provides an overlapping of neighbourhoods. Overlapping, though, introduces its own significant complexities. Instead, we propose constructing an image-wide context for the interpolation process, without the added complexity of overlapping blocks, by performing the interpolation in the domain of the DWT (Discrete Wavelet Transform).

For a DWT-based RS system, we consider only a single stage of 2D DWT, based on the Daubechies 9/7 biorthogonal wavelet kernel. In this case, $\mathbf{x}[\mathbf{n}]$ denotes a collection of four subband samples, one from each of the LL, LH, HL and HH subbands.

Method	Bilinear	Direct RS	LL only	DWT RS
PSNR (dB)	28.9	31.0	30.9	31.9

Table 1. DWT RS compared to other scaling methods.

In this way, we preserve the one-to-one correspondence between low resolution image neighbourhoods $\mathbf{z}[\mathbf{n}]$ and the high resolution synthesized output vectors $\mathbf{x}[\mathbf{n}]$. Apart from this change, we proceed exactly as before. In fact, the jump from the previous RS system to a DWT-based RS system involves only the training of a new set of MMSE synthesis parameters, $\mathbf{A}_j$ and $\boldsymbol{\beta}_j$. The classification parameters are unaffected. To interpolate a low resolution input image, we first synthesize the vectors, $\hat{\mathbf{x}}[\mathbf{n}]$, and then apply DWT synthesis to create the high resolution image from its subbands.

Early results point favourably to the improved performance of the DWT approach when compared to classical methods, Direct interpolation using RS and reconstruction from the low frequency approximation alone as shown in Table 1. It is anticipated that further refinements will address the current disparity between the modest improvement obtained and the computational and complexity costs necessary to achieve it compared to simpler methods.

6. CONCLUSIONS

In this paper we investigate a number of underlying hypotheses of recent work on image interpolation using resolution synthesis. We have proposed an alternate DWT-based resolution synthesis strategy, which leverages off the same Gaussian mixture model for low resolution image classification. We also introduced a number of algorithmic and computational refinements to improve the practical usefulness of the resolution synthesis approach to image interpolation. Evidently, there are several promising directions for significantly improving the performance and efficiency of the proposed framework.

7. REFERENCES

[1] S.-H. G. Chang, Z. Cvetković, and M. Vetterli, "Resolution enhancement of images using wavelet transform extrema extrapolation," *IEEE Trans. Acoust. Speech and Sig. Proc.*, vol. 4, pp. 2379–2382, May 1995.

[2] Z. Ye, Q. Guohui, and L. Shaobin, "Wavelet transform and multi-resolution synthesis of texture images," *Proc. IEEE Region 10 Conf. Computer, Communication, Control and Power Engineering*, pp. 442–445, 1993.

[3] C. Atkins, C. Bouman, and J. Allebach, "Optimal image scaling using pixel classification," *Proc. IEEE Int. Conf. Image Proc.*, vol. 3, pp. 864–867, Sep 2001.

[4] C. Atkins, "Classification-based methods in optimal image interpolation," *Ph.D. thesis, Purdue*, 1998.

[5] C. Atkins, C. Bouman, and J. Allebach, "Treebased resolution synthesis," *Proc. Image Quality, Image Capture Systems Conf., Savannah, GA*, pp. 405–410, Apr 1999.

[6] F. Rolf, "Single-link clustering algorithms," *P. R. Krishnaiah and L. N. Kanal, eds., Handbook of Statistics, North-Holland Publishing Company*, vol. 2, pp. 267–284, 1982.