Morphologically Unbiased Classifier Combination through Graphical PDF Correlation

David Windridge and Josef Kittler

Centre for Vision, Speech and Signal Processing
Dept. of Electronic & Electrical Engineering, University of Surrey
Guildford, GU2 5XH Surrey, United Kingdom
Telephone: +44 1483 876043
d.windridge@eim.surrey.ac.uk

Abstract. We reinterpret the morphologically unbiased 'tomographic' method of multiple classifier combination developed previously by the authors as a methodology for *graphical PDF correlation*. That is, the original procedure for eliminating what are effectively the back-projection artifacts implicit in any linear feature-space combination regime is shown to be replicable by a piecewise morphology matching process. Implementing this alternative methodology computationally permits a several orders-of-magnitude reduction in the complexity of the problem, such that the method falls within practical feasibility even for very high dimensionality problems, as well as resulting in a more intuitive description of the process in graphical terms.

1 Introduction

Within the field of machine learning there has been a considerable recent interest in the development of *Multiple Classifier Systems* [1-6], which seek to make use of the divergence of classifier design methodologies to limit *a priori* impositions on the morphology applicable to the decision boundary, such that a consistent boost in classification performance is observed.

In establishing a general theoretical framework for such approaches, the authors have determined previously [7-10] that classifier combination in virtually all of its variant forms has an aspect which may be regarded as an approximate attempt at the reconstruction of the combined pattern space by tomographic means, the feature selection process in this scenario constituting an implicit Radon integration along the lines of the physical processes involved in NMR scanning, etc (albeit of a multi-dimensional nature). It was thereby ascertained that an optimal strategy for classifier combination can be achieved by appropriately restructuring the feature-selection algorithm such that a fully-constituted tomographic combination (rather than its approximation) acts in its most appropriate domain: that is, when the combination is comprised of classifiers with *distinct* feature sets.

As in medical imaging, this fully constituted tomographic combination necessarily involves the application of a deconvolution procedure to a *back-projected*[1] space, which, in the context of pattern-recognition, we were able to demonstrate amounted to the composite probability density function (PDF) constructed implicitly by the Sum-Rule decision scheme [7]. In conventional implementations of tomography [eg 11], such deconvolution is most usually accomplished via a collective prior filtering of the Radon integrals. This would take the form of a differentiation operator that acts to remove what, in the reconstructive space, would (for the case of perfect angular sample coverage) amount to convolution by an $|1/r|$ function. The very low angular sampling implied by feature-selection, however, means that only the broadest-scale structure of the back-projection artifacts can be removed in this fashion, leaving a deconvolution with angular artifacts that are still overtly and unrepresentatively correlated with the feature axes - precisely the eventuality that we are seeking to eliminate, having tested for the possibility of *actual* correlation at an earlier stage of feature selection.

The most appropriate approach [8] to removing these spurious correlations is therefore that of unfiltered (post-)deconvolution, via an adaptation of a procedure developed for use with astrophysical data: namely, the Högbom deconvolution algorithm [12], which was specifically engineered for the removal of telescopically-correlated morphology. The iterative nature of this technique allows a piece-by-piece removal of systematic artifacts such that, in its unmodified and mathematically ideal form, the procedure can be considered to impose an *a priori* condition of least possible dependence of the recovered morphology on the individual classifiers' feature-axis geometry. Thus, the procedure embodies a distinct methodology for distinguishing between the degenerate solutions that all methods of deconvolution are required to address whenever there exist zeros in the Fourier transform of the entity to be deconvolved.

It is, however, possible to view the degenerate form that the Högbom methodology reduces to in the particular environment of the stochastic domain from an entirely different perspective: that of graphical PDF correlation. In setting out this relation precisely, we shall first seek to describe how the Högbom algorithm is implemented within the Sum-Rule domain for the two-dimensional case[2].

2 Nature of Högbom Deconvolution in the Two-Dimensional Sum-Rule Domain

It was was established in [7-10] that the back-projection artifact implied by the composition (via the sum rule) of two classifiers representing the differing decision spaces; x and y, is the equivalent of a "cross" of infinitesimal width (ie $P_{\text{artefact}}(x,y) = \delta(x) + \delta(y)$, with δ the Dirac delta function). It is consequently this spurious entity (modified appropriately to account for the discrete

---

[1] For definitions of this, and other tomographic terms, refer to [11]

[2] We shall retain the two-dimensional paradigm for simplicity throughout the paper, the generalisation to higher dimensionalities being self-evident unless otherwise stated (or else see [7-10] for a full description).
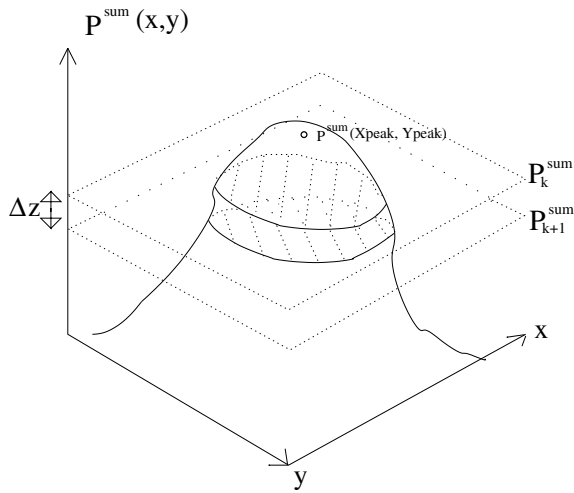
Fig. 1. The composite PDF in the Sum-Rule space

sampling of the PDF inherent in a computational methodology) that we are seeking to remove (deconvolve) from the composite feature-space PDF through recursive Högbom subtraction. In the two-dimensional case this is enacted as follows: a counter value, z, is set at the peak value of the Sum-Rule space, with a recursive scanning cycle then initiated to establish the set of all coordinates within a probability density value $| < \Delta z|$ below this. After registration of these points in a deconvolution matrix (so called because it will ultimately constitute the proposed deconvolution), through the addition of a value Δz to any existing value at the designated coordinates, a set of cross-artifacts centred on the corresponding points of the sum-rule space are then subtracted consecutively. This process is repeated until a subtraction is proposed by the procedure that would yield negative values in the Sum-Rule space, with, hence, a *complete* deconvolution resulting in a residual-free, zero-valued space. (Note that in the application to astronomical data, the procedure must rely instead on a stochastic criterion of completion in the absence of an absolute zero-point, namely the indistinguishability of the histogram of point values from a Poissonion "noise" distribution). The terminal point of the procedure therefore invariably represents (even in the absence of a proper termination) a positive-definite solution in the deconvolution matrix, as demanded by probability theory. This procedure will be more fully quantified at the computational level in the following section, however, we must first address a significant difficulty that arises with this approach:

2.1   Finite Δz Issues

It rapidly becomes apparent in any computational implementation of the Högbom deconvolution algorithm in the tomographic domain that the issue of the necessarily finite setting of the value Δz becomes non-trivial. It is intuitively

obvious that the process achieves mathematical ideality only in the asymptotic limit: $\Delta z \to 0$, in which case each iterative stage registers an unambiguous set of discrete points at uniform height. However, the fact that any computational implementation must rely on a finite value of Δz gives rise to complications that have consequences that go far beyond issues of sampling accuracy: selecting different values of Δz for the situation set out above would in fact generate a vastly divergent set of completions at the termination of the procedure. Mitigating this consideration, however, is the fact that these terminal sets *do* represent consistent deconvolutions given the initial data, in the sense that the recovered distributions all revert, if re-convolved by the cross-artifact, to the originally specified Sum-Rule space.

It would perhaps, therefore, seem logical to choose $\Delta z = 0$ as being in some sense a favoured option on (as yet not fully specified) *a priori* grounds. However, any practical implementation must take place within a discretely-sampled computational setting: in proposing a finite Δz procedure that does not experience the above problem (ie, whose solution has no explicit dependence upon the value of Δz), we have to consider more systematically what is taking place during the *simultaneous* subtraction of cross-artifacts implicit in each iteration. As is uniquely the case for tomographic reconstruction of a pattern-space, these subtraction entities share an identity with the form of the axial system (that is to say, constitute a complex of intersecting quadrilaterals of varying dimensionality [8]). We can therefore appreciate that the simultaneity of the subtraction immediately gives rise to an irreconcilable ambiguity: we see that the *overlap* of these entities necessarily gives rise to further intersections at specific points of the pattern space, the artifacts around which *are of the same form as the axial system*, which are hence not in any real sense distinguishable from the original points at which axial artifacts are subtracted.

It is therefore apposite to propose as a modification of the Högbom method (when acting in the expert fusion domain), that these additional points are put forward as candidates for registration *alongside* the originals. It is, in a sense, therefore possible to regard this modification as summing over all possible deconvolution solutions that we earlier encountered at the iterative level. This amounts to applying the most conservative criterion of PDF correlation *within* the terms of the Högbom approach, while maintaining the most presumptive *a priori* condition on the feature correlation in more general terms (which is to say, imposing an assumption of minimal feature dependence on the axes, the alternative having been eliminated at the feature selection stage).

In visualising this alternative approach, it is most useful to focus on the effect that the Högbom algorithm has on the PDFs *constituting* the combination, rather than the Sum-Rule space, as we have hitherto done. The nature of the Högbom iteration is also rendered far more graphically evident from such a perspective:

2.2   PDF-Centred Approach

As we have thus far understood it, then, the commencement of the Högbom procedure consists the determination of the peak position of the Sum-Rule space
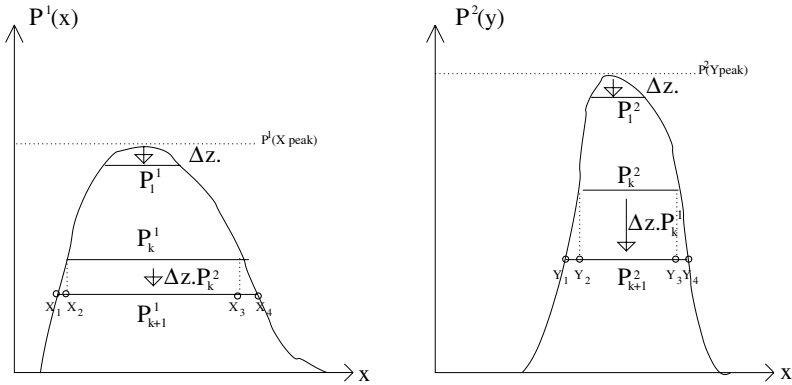
Fig. 2. Requisite subtractions from the two PDFs constituting the combination in the modified methodology: note the presence of P_k^2 in the first diagram's subtraction (and vice-versa)

($P^{sum}(X_{peak}, Y_{peak})$ from fig. 1), and the derivation of the set of points, P_1^{sum}, that lie in the probability density range ($P^{sum}(X_{peak}, Y_{peak}) \rightarrow P^{sum}(X_{peak}, Y_{peak}) - \Delta z$), prior to subtracting a series of cross-artifacts centred on those points. We should now like to associate these points with particular sets of ordinates in the PDF domain such that it is possible to view the 3-dimensional process of fig. 1 within the 2-dimensional format of fig. 2. This would not in general be possible to do in a straightforward fashion if the subtraction entity were of an arbitrary form. However, the fact that the subtraction artifact mirrors the axial system means that it may be equivalently represented as the *independent* summation of 1-dimensional Dirac delta functions (convolved by the sampling element Δx) centred on the appropriate ordinates of the *PDF domain*. The process of subtraction of a *single* artifact in this domain therefore acquires the intuitive aspect of a subtraction of individual delta functions from the appropriate points of the respective classifier PDFs ($\delta(x - x_0)\Delta x$ from $P^1(x)$, and $\delta(y - y_0)\Delta x$ from $P^2(y)$, in our case).

Although this situation readily generalises to the arbitrarily-dimensioned case, it becomes somewhat more complex for multiple subtractions of the type indicated earlier, in that the subtraction of cross-artifacts centred on the additional set of points created by the intersections of the artifacts (arising from the originally detected points) leads to an asymmetry in the corresponding PDF domain subtractions: the particular value to be subtracted from each of the ordinals in a particular PDF turns out to require a proportionality to the subtractions in the remaining PDFs constituting the combination. This is illustrated in fig. 2 for a mid-point of the deconvolution's execution (since we are required to externally impose an infinitesimal subtraction on the *first* iteration of the sequence $k = 1, 2 \ldots$, which cannot, therefore, exhibit this effect explicitly).

A subtraction, then, of the points above P_{k+1}^{sum} (points above P_k^{sum} having been assumed to have been removed by previous iterations) leads to a replacement of the ordinal sets: $\{x|P^1(x) = P_k^1\}$ with $\{x|P^1(x) = P_{k+1}^1\}$ and $\{y|P^2(y) = P_k^2\}$ with $\{y|P^2(y) = P_{k+1}^2\}$: that is to say, a reduction of $\Delta z|P_k^2|$ and $\Delta z|P_k^1|$ in $P_k^1(x)$ and $P_k^2(y)$, respectively (with a corresponding registration of Δz in the deconvolution matrix for the coordinate-set $\{x|X_1 \leq x \leq X_2\} \times \{y|Y_1 \leq y \leq Y_2\}$, that is, all combinations of ordinals over this range). Note in particular the transfer of width information from one PDF to the other, giving rise to the mutually morphologically dependent convergence alluded to earlier: we are then now implicitly regarding the PDFs, not as maps $\mathcal{R} \to \mathcal{R}$, but rather as morphological entities delineating 'areas' in an ordinate-probability space.

The fact that these points lie in bands is critical to the method's economy, and a consequence both of the explicit inclusion of the intersection point-sets (of which more later), but also of the particular nature of this stage-by-stage remapping. For the set of ordinates newly incorporated into the $(k+1)$'th iteration to be consistent with the line defined by the ordinate set arising from the k'th iteration, this involves imposing a transformation:

$$\{P_x^1\} \to \{P_{X_1}^1\} \quad \forall \, (X_1 < x < X_2) \text{ and } \forall \, (X_3 < x < X_4) \tag{1}$$

$$\{P_y^2\} \to \{P_{Y_1}^2\} \quad \forall \, (Y_1 < y < Y_2) \text{ and } \forall \, (Y_3 < y < Y_4), \tag{2}$$

at each new stage of the process, such that each new ordinate set is contained within its predecessor. Thus, the algorithmic recursion applies solely now to these ordinal sets (two single-dimensioned entities, rather than to a single Sum-Rule density function of three dimensions). It should also be noted that this approach is equally valid for the more complex case of multiply-peaked PDFs, the extension to the mapping protocol being a matter of straightforward extrapolation.

The other issue which we have yet to approach systematically within this framework arises in relation to multiple subtractions, and concerns the aforementioned ambiguity arising from the cross-correlation between subtractive entities. In fact, it transpires that a quantitative treatment of this effect is rendered significantly more straightforward on consideration within the PDF domain: in removing multiple delta-function elements from the individual density functions, all of the interstitial "overlap" artifacts are implicitly dealt with at the same time. This can be illustrated in the two-dimensional case via an appreciation of the fact that the subtraction of delta-function elements centred on the P^1 ordinals; X_1 and X_2, and the P^2 ordinals; Y_1 and Y_1, would imply a subtraction of cross artifacts centred on; $P^{sum}(X_1, Y_1)$, $P^{sum}(X_2, Y_1)$, $P^{sum}(X_2, Y_2)$ and $P^{sum}(X_1, Y_2)$: that is to say, the complete set of detectable points in the Sum-Rule domain *as well as* their subtraction-artifact overlaps. The only remaining issue to address in relation to the PDF-centred approach to Högbom deconvolution is then the construction of the actual *co*-ordinates for registration in the deconvolution matrix, which, it is readily seen, are just the set of all permutations of detected ordinals within the current iteration.

In this manner, by switching to a PDF-oriented approach, necessitating what is effectively a varying Δz methodology within which the issue of multiple reg-

istrations and subtractions is dealt with *automatically*, we have effectively dissolved the distinction between PDF point-detection, artifact-correlation and artifact subtraction, generating a significant speed increase through the fusion of the three space-scanning processes implicit in the tomographic method, as well as a further, arbitrarily large speed increase determined by the implicit fusion of the Δz parameter with the morphology of the PDFs (through the inclusion of cross-sectional magnitude terms *within* each iteration).

We shall determine more precisely the effect that this has on the computational efficiency of the tomographic method as follows:

2.3   Computational Implementation

The first economization attributable to the new approach, arising as a consequence of the implicit identification of the peak-search, peak-correlation and artifact-subtraction procedures, reduces a process of originally $\sim [X]^{2n}[X^{n-1} + X]$ cycles to around X^{n-1} cycles (n being the dimension of the reconstructive space, and X its linear sampling resolution: the square brackets denote a maximum value). This is determined in the following way: within the unmodified Högbom procedure each iterative scan of the Sum-Rule space to obtain a set of points for subtraction carries with it a penalty of X^n cycles. Because Δz is not correlated with the PDF cross-sections as it is in the modified case, the requisite analysis of subtraction-artifact overlapping will require that the additional interstitial points are all individually constructed and registered within the deconvolution matrix. In the worst case scenario, when the ordinates of the detected points cover the entirety of the feature axes, this would amount to an implicit scan over the entire reconstructive space, requiring an additional computation of $[X]^n$ cycles (a scan being effectively the exhaustive cyclic construction of ordinal permutations).

A deconvolution-artifact subtraction at each of these points would then require a further scanning agent to act over the reconstructive space, ostensibly involving a further X^n cycles *per point*. However, it is possible to break the artifacts down into their constituent iterations to obtain a reduction in this. That is, if the set of classifiers constituting the combination have an individual feature-dimensionality given by d_i, then this would represent a required per-point cycle count of magnitude $(X^{d_1} + X^{d_2} + X^{d_3} \ldots)$ in order to perform the subtraction. In execution terms, this represents a *maximum* of $X^{n-1} + X$ cycles (the best case scenario being just nX cycles, or $2X$ in our example). The total cycle count per iteration for the Högbom method is therefore: $X^n[X]^n[X^{n-1}+X]$, where it is understood that this (and all following terms) represent worst-case scenarios.

By contrast, the proposed alternative, in combining the detection, correlation and subtraction procedures, permits a cycle count of only X^n per iteration. This comes about through combining the activity of a detection/subtraction scan that acts over just the constituent PDF feature dimensions (which would in itself now carry only a $[X^{n-1} + X]$ cycle penalty) with a correlation analysis (which would normally constitute an *additional* $[X]^{d_1} + [X]^{d_2} + [X]^{d_3} \ldots = [X]^n$ cycles per point), such that the correlation analysis, in generating every possibly

ordinal permutation, now implicitly performs both the detection and subtraction operations in the manner described above.

It is possible, within the proposed alternative to tomographic combination, to further improve on this performance for the particular case of the constituent classifiers constituting point-wise continuous PDFs, through the introduction of a second-order computational economy. We note in fig. 2 that P_k^1 is fully contained within the set P_{k+1}^1, with only the sets $P^1(X_1) \to P^1(X_2)$ and $P^1(Y_1) \to P^1(Y_2)$ then contributing a new behavioural aspect to the $(k+1)$'th iteration (and similarly for $P^2(Y)$). Thus, the newly correlated and registered points in the $(k+1)$'th iteration will all lie <u>inside</u> of the P^{sum} region defined by the coordinate range: $(X_1 \to X_4, Y_1 \to Y_4)$, and <u>outside</u> of the smaller region $(X_2 \to X_3, Y_2 \to Y_3)$. Hence (and this is equally true for multiply-peaked PDFs), it becomes possible to simply discard this region within the correlation analysis (by far the most computationally expensive part of the proposed methodology), leaving only the originally specified artifact subtraction to perform, at a penalty of $[X^{n-1} + X]$ cycles. In algebraic terms this results in a cycle count reduction to:

$$\{[X^{n-1} + X]\} + \{(X + dx)^n - X^n\} \approx \{[X^{n-1} + X]\} + \{n\, dx\, X^{n-1}\} \quad (3)$$

(the later bracketed term in the addition constituting the generalisation of the above reasoning to arbitrary dimensionality, and dx being the sampling element [of similar fractional width to Δz]). This is clearly, then, a very substantial additional saving.

As a final note, it is evident that the number of iterations is itself a key dictator of execution time and, as we have observed, is a quantity that need not necessarily be fixed, a fact from which we have considerably benefited. However, the actual *value* of the number of iterations is governed by PDF morphology, and consequently not straightforwardly enumerable. The original Högbom method, however, does not suffer this limitation, requiring a fixed $(P_{max}^1 + P_{max}^2 + \ldots)/\Delta z$ iterations to execute, and serves as an upper limit for the modified procedure (although in practice we would expect the actual value to be a small fraction of this).

Thus, in the final analysis, the total cycle count for the more efficient methodology can be written:

$$[(P_{max}^1 + P_{max}^2 + \ldots)/\Delta z]\{[X^{n-1} + X]\} + \{ndxX^{n-1}\}$$
$$\approx [(P_{max}^1 + P_{max}^2 + \ldots)]\{[X^{n-1} + X]/\Delta z\} + \{nX^{n-1}\} \quad (4)$$

as opposed to:

$$(P_{max}^1 + P_{max}^2 + \ldots)/\Delta z\{X^n[X]^n[X^{n-1} + X]\} \quad (5)$$

under the original proposal.

3   Conclusion

We have set out to reinterpret the tomographic method of classifier combination within its most natural context, significantly reducing the computation time

involved to the extent that the method now poses very little obstacle to practical implementation.

The basis of this efficiency gain is the observation that, viewed in terms of the constituent PDFs, the three chief computational components of the recursive tomographic procedure (the peak-seek, the peak correlation analysis and the subtraction/registration of correlated components) need not actually be performed on an individual basis, reducing an iteration requirement of $X^n[X]^n[X^{n-1} + X]$ computational cycles to a maximum of X^n, with the further possibility of an order of magnitude decrease in this figure for point-wise continuous classifiers. Finally, there are further (if not precisely quantifiable) gains arising from dynamically varying the Δz parameter throughout the procedure.

References

1. R. A. Jacobs, "Methods for combining experts' probability assessments", Neural Computation, 3, pp. 79-87, 1991
2. J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 3, 1998, 226-239
3. L. Lam and C. Y. Suen, "Optimal combinations of pattern classifiers", Pattern Recognition Letters, vol. 16, no. 9, 1995, 945-954.
4. A. F. R. Rahman and M C Fairhurst, "An evaluation of multi-expert configurations for the recognition of handwritten numerals", Pattern Recognition Letters, 31, pp. 1255-1273, 1998
5. A. F. R. Rahman and M C Fairhurst, "A new hybrid approach in combining multiple experts to recognise handwritten numerals", Pattern Recognition Letters, 18, pp. 781-790, 1997
6. K. Woods, W. P. Kegelmeyer and K Bowyer, "Combination of multiple classifiers using local accuracy estimates", IEEE Trans. Pattern Analysis and Machine Intelligence, 19, pp. 405-410, 1997
7. D. Windridge, J. Kittler, "An Optimal Strategy for Classifier Combination: Part 1: Multiple Expert Fusion as a Tomographic Process", (PAMI, Submitted)
8. D. Windridge, J. Kittler, "An Optimal Strategy for Classifier Combination: Part 2: General Application of the Tomographic Procedure", (PAMI, Submitted)
9. D. Windridge, J. Kittler, "Classifier Combination as a Tomographic Process", (Multiple Classifier Systems, LNCS. Vol. 2096 , 2001.)
10. D. Windridge, J. Kittler, "A Generalised Solution to the Problem of Multiple Expert Fusion.", (Univ. of Surrey Technical Report: VSSP-TR-5/2000)
11. F. Natterer, Proceedings "State of the Art in Numerical Analysis", York, April1-4, 1996.
12. J. Högbom, "Aperture synthesis with a non-regular distribution of interferometer baselines", Astrophys. J. Suppl. Ser., 15, 417-426, 1974