COARSE-LEVEL OBJECT RECOGNITION USING INTERLEVEL PRODUCTS OF COMPLEX WAVELETS

Ryan Anderson, Nick Kingsbury, Julien Fauqueur

Signal Processing Group, Dept. of Engineering, University of Cambridge, UK

ABSTRACT

This paper introduces the Interlevel Product (ILP) which is a transform based upon the Dual-Tree Complex Wavelet. Coefficients of the ILP have complex values whose magnitudes indicate the amplitude of multilevel features, and whose phases indicate the nature of these features (e.g. ridges vs. edges). In particular, the phases of ILP coefficients are approximately invariant to small shifts in the original images. We accordingly introduce this transform as a solution to coarse scale template matching, where alignment concerns between decimation of a target and decimation of a larger search image can be mitigated, and computational efficiency can be maintained. Furthermore, template matching with ILP coefficients can provide several intuitive "near-matches" that may be of interest in image retrieval applications.

1. INTRODUCTION

This paper considers the problem of efficient recognition of objects in images. We introduce a new method of processing the outputs of a directional multiscale transform so as to provide approximate invariance to local shifts of the image data while retaining a strong ability to distinguish different patterns within the pixels of a local region. Our method is based on the Dual-Tree Complex Wavelet transform (DT CWT) [1] and we use the name InterLevel Product (ILP) to describe the subsequent processing of the complex DT CWT coefficients. The phases of the ILP coefficients indicate the type of features present at each scale and subband orientation, and the ILP magnitudes are proportional to the magnitudes (importance) of these features at two adjacent scales. These coefficients exhibit strong invariance to shifts of the decimation grid, and therefore provide a rich, reliable representation of the original image at multiple scales.

With this new tool, we demonstrate multiscale measures that can simultaneously accelerate template-matching methods for 2-D object recognition and increase the information available to evaluate near-matches. This report concentrates mainly on introducing the ILP, describing its relationship to semantic image content, and showing its ability to successfully match images at coarse scales. Alternate coarse-level matching methods can be found in [2] and [3].

This work has been carried out with the support of the UK Data & Information Fusion Defence Technology Centre.

2. MULTISCALE MATCHING METHODS

Classic template matching with a normalized cross-covariance (NCC) is performed pixel-by-pixel with the calculation below:

$$\gamma(x,y) = \frac{\sum_{\alpha,\beta}[S(\alpha-x,\beta-y)-\bar{S}_{\alpha,\beta}][T(\alpha,\beta)-\bar{T}]}{\sqrt{\sum_{\alpha,\beta}[S(\alpha-x,\beta-y)-\bar{S}_{\alpha,\beta}]^2 \sum_{\alpha,\beta}[T(\alpha,\beta)-\bar{T}]^2}} \quad (1)$$

$\gamma(x,y)$ is the correlation value between an $N \times M$ target image, T, and equivalently sized candidate regions, centered at (x,y) of a (typically) much larger $X \times Y$ search database image S. The location of the best match candidate is found at $(x,y) = \max_{(x,y)\in(X,Y)} \gamma(x,y)$, and $\gamma(x,y)$ itself indicates the strength of the match, with $\gamma(x,y) = 1$ indicating a perfect match.

The NCC method has several disadvantages of note. First, its computational complexity is high, due to the need for pixel-by-pixel comparison at each (x,y) offset of the candidate image. Several methods have been introduced that successfully reduce these computations, such as normalization after matching in Fourier domain [4]. A second disadvantage of the NCC is that the correlation measure, $\gamma(x,y)$, is a simple scalar value that does not provide significant insight into the nature by which the target T and candidate region $S(x,y)$ differ.

Multiscale template matches have the potential to solve the two disadvantages highlighted above. Computation can be reduced by first matching coarse-level, decimated representations of the target and the search database. Each decimation by 2:1 reduces the block size by 4:1 and the number of search locations also by 4:1. Hence we obtain a 16:1 reduction in the computation for each level of decimation.

The largest difficulty with multiscale template matching is that, in the general case, a decimated representation of the target will be based upon a reference grid that is different from that of the search image. There is potential for misalignment of up to 75%, where the closest matching search

coefficient for a given target coefficient is calculated from only 25% of the same pixel values. For example, Figure 1 shows a target and search image, each decimated by 3 levels, so that an 8×8 $(2^3 \times 2^3)$ patch of image is represented by a single coefficient. In this case, the candidate subimages (shown by circles) could be misaligned by up to 4 pixels vertically *and* 4 pixels horizontally from the target subimages (shown by triangles). Hence each triangle coefficient in the target would be calculated from no more than 25% of the pixels which contribute to each of the four surrounding circle coefficients in the candidate image. To

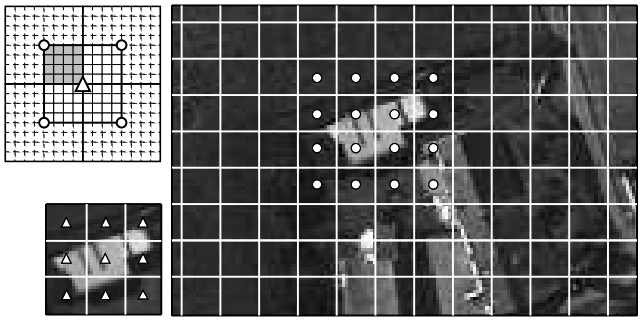

Fig. 1. Example of misalignment in decimated multiscale template matching.

compensate for this potential error, one could blur the original image prior to decimation in order to extend the overlap of image content into several coefficients; but then critical details, particularly edge information, will be lost. The Discrete Wavelet Transform (DWT) provides a good basis for preserving relevant edge information at a given decimation level; however, the DWT is highly shift-dependent, and will still therefore suffer from reference grid misalignment. We therefore consider complex wavelets, specifically the Dual-Tree Complex Wavelet Transform (DT CWT) introduced in [1]. The coefficients of the DT CWT accurately reflect local edge content; the magnitudes are relatively shift invariant, and phases change linearly with the offset of a local edge relative to the coefficient. The simplest way to use the DT CWT in a decimated template match is to match the magnitudes of the coefficients (a similar approach can be found in [5] for stereo matching). As can be seen in section 5, this method shows distinct advantages to other methods of the same resolution.

DT CWT coefficient magnitudes are shift-invariant, but they tell us little about the structure of the underlying image. For instance, it is difficult to distinguish a ridge from a step edge using just DT CWT coefficient magnitudes at a coarse scale. Such ambiguities increase the probability of false matches at coarse scales, and so increase the search time for coarse-to-fine matching. We therefore look at the relative phases of DT CWT coefficients across scale to provide us with such structural information.

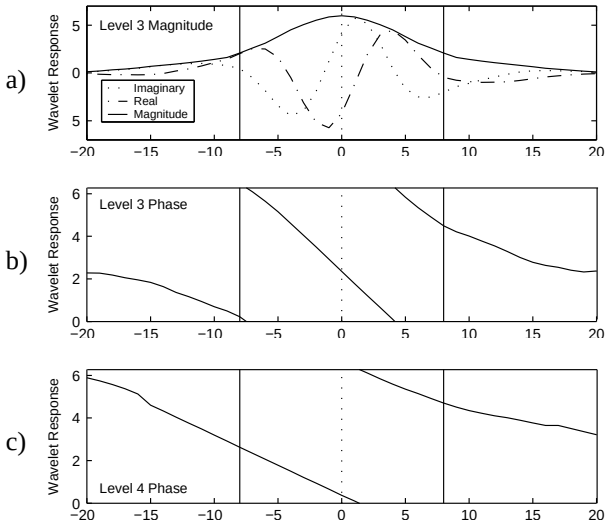

Fig. 2. Magnitude (a) and phase (b) of a DT CWT coefficient at level 3 in the presence of a step edge at pixel offset x, and the equivalent phase of a level 4 coefficient (c). Vertical lines indicate the position of the central level 3 DT CWT coefficient and its adjacent neighbours at a distance of 2^3=8 pixels.

3. THE INTERLEVEL PRODUCT

If we observe a 1-D level-3 DT CWT coefficient while a step edge is translated past it, its magnitude and phase behave as shown in Figure 2. The horizontal axes show the translation x and fig 2(a) shows the complex components of the coefficient magnitudes while fig 2(b) shows the phase. Note that the phase of the coefficient θ_3 changes approximately linearly with x, as long as the step edge is within the inter-coefficient distance, which is 2^3 pixels for level 3 and 2^l for level l in general.

The same step edge "wave" is also observable at level 4 - as it is a multiscale edge present in both scales - so the parent of the original level 3 coefficient will undergo rotation by θ_4; however, the rotation will occur at half the speed of θ_3. Hence the spatial rotation rates are related by $\frac{d\theta_3}{dx} = 2\frac{d\theta_4}{dx}$.

The constancy of this ratio allows us to create a coarse-level shift-invariant complex feature, the InterLevel Product (ILP), whose phase is $\theta_\Delta = \theta_3 - 2\theta_4$ and whose magnitude is the product of the level-3 and level-4 coefficient magnitudes. This can easily be achieved by taking the product of each level-3 coefficient with the complex conjugate of a corresponding level-4 coefficient whose phase has been doubled. Figure 3 shows this process. Each of the 16 columns in the whole figure shows the coefficients corresponding to 16 different shifts of an input step, as in (a). (b) and (c) show the level-4 and level-3 coefficients, while (d) shows the result of bandpass interpolating the level-4 coefficients at the

level-3 sample locations and then phase-doubling them. Finally (e) shows the result of multiplying the coefficients in (c) by the conjugates of those in (d) to give the ILP coefficients. Note the almost constant (shift-invariant) phases of the larger ILP coefficients and the very slow variation of their magnitudes with shift. The phase of an ILP coefficient, θ_{Δ}, is dependent mainly upon the nature of the dominant multiscale feature in its vicinity (see Table 1).

This operation can be applied to the six individual subbands of a 2-dimensional DT CWT at each level, to represent edges and ridges in each direction of a 2-D image. In this case, ILP coefficients describe 2-D edges or edges near the orientation of each subband. Calculation of the 2-D ILP values is equivalent to the 1-D ILP, although the bandpass interpolation step in (d) involves a 2-D frequency shift to the origin of the frequency plane, a lowpass interpolator, and a reversal of the frequency shift.

4. ILP OBJECT RECOGNITION

As described above, we first transform the $N_0 \times M_0$ target T to a $N_l \times M_l \times 6$ complex ILP representation $\chi_l^{(T)}$, and the $X \times Y$ candidate image S to a $X_l \times Y_l \times 6$ representation $\chi_l^{(S)}$. As $N_l = \frac{N_0}{2^l}$, $M_l = \frac{M_0}{2^l}$, $X_l = \frac{X_0}{2^l}$, and $Y_l = \frac{Y_0}{2^l}$, this operation greatly reduces the size of the compared datasets when the decimation level $l > 2$.

We then wish to find areas of the candidate image whose ILP phases match the ILP phases of our target at points of strong saliency. Thus, we simply multiply each ILP coefficient in the target $\chi_l^{(T)}$ by the complex conjugate of corresponding ILP coefficients from an equivalently sized, decimated, candidate region of S, $\chi_l^{(S)}(i,j)$, as below:

$$r_{(i,j,\alpha,\beta,d)(l)} = \left[\chi_l^{(S)}(\alpha - i, \beta - j, d)\right]^* \times \chi_l^{(T)}(\alpha, \beta, d)$$

In our models, (i,j) represents the top left corner of the region of comparison in the search image ILP; $\alpha = 1...N_l$, $\beta = 1...M_l$ are the non-negative integers representing each spatial coefficient location; and $d = 1...6$ is the directional subband. At each coefficient location, the result $r_{(i,j,\alpha,\beta,d)(l)}$ is a complex value that will be closely aligned with the positive real axis, in the case of a match, or of random phase otherwise. Where the aligned coefficients are of large magnitude (strongly salient) the result will be a large positive real number. Accordingly, a summation of the real components of these r values will give us a correlation measure for the match. We also normalize this sum by the magnitudes of the ILP coefficients, in a manner analogous to the NCC:

$$R_{(i,j)(l)} = \Re\left[\frac{\sum_{\alpha,\beta,d} r_{(i,j,\alpha,\beta,d)(l)}}{\sqrt{(\sum_{\alpha,\beta,d} |\chi_l^{(S)}(\alpha-i,\beta-j,d)|^2) \times (\sum_{\alpha,\beta,d} |\chi_l^{(T)}(\alpha,\beta,d)|^2)}}\right] \quad (2)$$

The estimated location of the upper left corner of the target at level l is $(\hat{i}_l, \hat{j}_l) = \arg\max_{(i,j)} R_{(i,j)(l)}$ in ILP

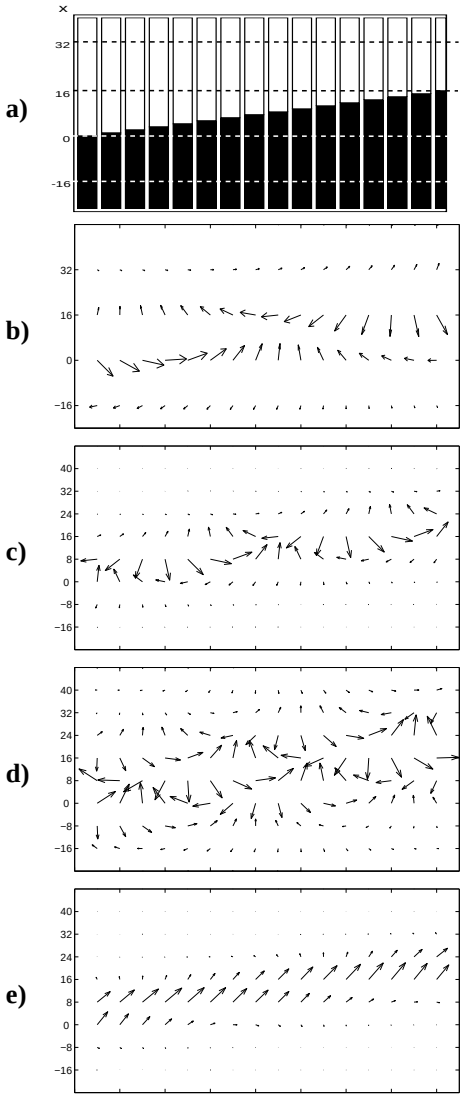

Fig. 3. 1-D Interlevel Product Calculation: a) 16 shifts of a step edge (horizontal lines show level 4 wavelet locations); b) level-4 DT CWT coefficients for each shift position; c) level-3 DT CWT coefficients; d) bandpass interpolated and phase-doubled level-4 coefficients; e) ILP coefficients.

Feature	ILP Angle
Positive Step	$45°$
Positive Impulse	$135°$
Negative Step	$-135°$
Negative Impulse	$-45°$

Table 1. Relationship between 1-D feature and complex phase of ILP coefficients.

coefficients, and is consequently $(\hat{x}, \hat{y}) = (\hat{i}_l \times 2^l, \hat{j}_l \times 2^l)$ in the original pixels.

5. ILP TEMPLATE MATCHING RESULTS

For testing, our search targets are 64×64 pixel images, selected from seven 4000×4000 pixel search images (Obtained from the "Window on the UK 2000" CD, available at `http://www.bnsc.org/wouk/wouk1.htm`). 200 of these targets contain buildings, with strong edges and correspondingly salient ILP coefficients, and the remaining 200 targets contain forest and vegetation.

For each of 400 target searches, we perform a comparison of coarse level $l = 3$ matches using the following decimation operators on the target, T, as well as the search database S (the contributing image):

- **ILP.** $\chi_l(i, j, d), d = 1...6$, as calculated in section 3, from the DT CWT of image T.
- **DT CWT Magnitudes.** $|W_l(i, j, d)|, d = 1 \ldots 6$, where W is the DT CWT of T.
- **Blurred and Decimated.** $\mathbf{B}_l(i, j); \mathbf{B} = \mathbf{A}_l^T \mathbf{T} \mathbf{A}_l$; $\mathbf{A}_l = \mathbf{I}_{N_l \times M_l} \bigotimes \mathbf{1}_{1 \times m}$, where $\bigotimes$ is the Kronecker product and $\mathbf{I}$ is the identity matrix. (from [2])
- **DWT Magnitudes.** $|D_l(i, j, d)|, d = 1 \ldots 3$, where D is the DWT of T.

Approaches to date ([2], [3]) have not addressed the specific effect of misalignment, and may assume that the decimation operator is aligned in both the target and search image. For our testing, the search image is deliberately offset by four pixels in each direction, creating the worst-case misalignment situation shown in Figure 1. We then calculate $R_{(i,j)}(l)$ and a best estimate $(\hat{x}, \hat{y})$ as described above.

In Table 2, we show the following values:

- "First Match" indicates the percentage of the 400 match attempts in which the estimated location was correct. This value should be as large as possible.
- $P_3(k)$ refers to the number of candidates in a subset of locations that is expected to contain the correct match, with k=98% and k=90% confidence respectively. This value should be as small as possible.

The ability of the ILP to estimate the correct match in the first instance is notably superior to other methods; note that, in particular, the DWT operator completely fails to cope with misalignment due to its excessive shift variance. For our test data of typical urban and vegetation images, 93% of the ILP match attempts at coarse-level were successful with a single match. It is only when one attempts to set the probability of a correct match to $k = 98\%$ that the P_3 window of candidates increases greatly. This increase occurs because certain objects - such as sections of a long empty road or fine-textured vegetation - require finer minutiae to distinguish an actual target from a similar patch. And, indeed, an operator may wish to preserve these near-matches for legitimate consideration as alternative targets, depending upon the application.

Decimation Operator	**First Match**	$\mathbf{P_3(k)}$ (k=98%)	$\mathbf{P_3(k)}$ (k=90%)
ILP	93%	156	1
DT CWT Magnitudes	71%	717	14
Blurred and Decimated	16%	5050	1265
DWT Magnitudes	0%	$> 10^6$	$> 10^6$

Table 2. Summary of the relative abilities of different decimation operators to prune candidate regions at level 3.

6. CONCLUSIONS AND FUTURE WORK

Our ILP transform has the ability to characterize an object's coarse-level features (ridges and edges) consistently with parsimonious, decimated coefficients. These coefficients are strongly invariant with respect to the alignment of the decimation operation. We have demonstrated high performance at target matching to illustrate these properties and the accelerative potential of performing recognition tasks with the ILP. We intend to develop a metric by which the ILP magnitudes of a target at each level dictates the coarsest level at which one can efficiently begin a target match. We are also investigating the relationship between ILP phase and image content further, and developing rotation- and scale-invariant models of image objects.

7. REFERENCES

[1] N.G. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals," *Journal of Applied and Computational Harmonic Analysis*, vol. 10, no. 3, pp. 234–253, 2001.

[2] M. Gharavi-Alkhansari, "A fast globally optimal algorithm for template matching using low-resolution pruning," *IEEE Transactions on Image Processing*, vol. 10, no. 4, pp. 526–533, April 2001.

[3] Sumit Basu, "Efficient multiscale template matching with orthogonal wavelet decompositions.," Tech. Rep., MIT Media Lab, May 1997.

[4] J. Lewis, "Fast normalized cross-correlation," *Vision Interface*, pp. 120–123, 1995.

[5] Fangmin Shi, Neil Rothwell Hughes, and Geoff Roberts, "SSD matching using shift-invariant wavelet transform.," in *British Machine Vision Conference*, September 2001.