

MULTISCALE KEYPOINT DETECTION USING THE DUAL-TREE COMPLEX WAVELET TRANSFORM

Julien Fauqueur, Nick Kingsbury and Ryan Anderson

Signal Processing Group, Department of Engineering,
University of Cambridge, United Kingdom
{jf330,ngk,raa37}@cam.ac.uk

ABSTRACT

We present a novel approach to detecting multiscale keypoints using the Dual Tree Complex Wavelet Transform (DTCWT). We show that it is a well-suited basis for this problem as it is directionally selective, smoothly shift invariant, optimally decimated at coarse scales and invertible (no loss of information). Our detection scheme is fast because of the decimated nature of the DTCWT and yet provides accurate and robust keypoint localisation, thanks to the use of the “accumulated energy map”. The regularity of this map is used to introduce a new mechanism for robust keypoint scale selection. Keypoints of different nature and size can be detected with limited redundancy, in a way which is consistent with our visual perception. Furthermore results show better robustness against rotation compared to the SIFT detector.

1. INTRODUCTION

We are interested in the problem of keypoint detection in images. By keypoints, we mean typically blobs, corners and junctions. These features have been also referred to as interest or salient points in the literature. In human vision, these localised features, along with edges, are perceived as privileged cues for recognising shapes and objects and are widely used in computer vision for various applications including object recognition, stereo matching, content-based image retrieval, mosaicing, motion tracking.

Various methods have been proposed for keypoint detection. The Harris corner detector [1] is not designed to be multiscale. Differences of Gaussian or “DoG” (as used in SIFT [2]) act as isotropic filters and therefore require an extra step to distinguish between keypoints and edges.

Wavelet theory provides a powerful framework to decompose images into different scales and orientations, which is coherent with the human perception. Wavelet transforms which are non-redundant (i.e. which produce no more coefficients than the original number of pixels) and invertible (i.e. the

exact image content can be reproduced just from the wavelet coefficients), such as the Discrete Wavelet Transform (DWT), proved to be very powerful for image compression as in JPEG 2000.

The DWT has been used by Loupas et al [3] for salient point extraction. However, the DWT is not robust to shift and is poorly directionally selective [4], which is a major obstacle for keypoint detection. Gabor wavelets have been designed to be directionally selective. They are robust to shift, since they are non-decimated, but they are therefore highly over-complete and hence highly computationally expensive. More recently, the DTCWT [4] was proposed and was shown to be a particularly suitable tool for image analysis as it is directionally selective (see figure 1), approximately shift invariant and has limited redundancy. We will show that these unique advantages over other multiscale decompositions make the DTCWT an ideal candidate for *multiscale, robust and computationally efficient keypoint detection*, which are desirable properties for visual recognition tasks.

In our approach, the keypoint energies measured from the decimated DTCWT coefficients at different scales are accumulated into a single smooth energy map. This “accumulated map” plays a key role since its peaks define the keypoint locations and its gradient is used to derive the keypoint scales.

In section 2, we introduce the wavelet-based keypoint energy measure. The generation of the accumulated map from multiscale keypoints energies and the keypoint localisation scheme are explained in section 3. The scale selection scheme is presented in section 4. In section 5, we will compare our approach with the SIFT detector according to their robustness to rotation and noise transformations. We will conclude in section 6.

2. DETERMINING THE KEYPOINT ENERGIES FROM THE DTCWT COEFFICIENTS

The DTCWT decomposition of an $w \times h$ image results in a decimated dyadic decomposition into $s = 1, \dots, m$ scales, where each scale is of size $w/2^s \times h/2^s$. At each decimated location of each scale, we have a set C of 6 com-

This work has been carried out with the support of the UK Data and Information Fusion Defence Technology Centre. The authors would like to thank James J. Ng for providing the code for bivariate shrinkage denoising.

plex coefficients, denoted as $C = \{\rho_1 e^{i\theta_1}, \dots, \rho_6 e^{i\theta_6}\}$, corresponding to responses to the 6 subband orientations, namely: $15^\circ, 45^\circ, 75^\circ, 105^\circ, 135^\circ, 165^\circ$. The directional information

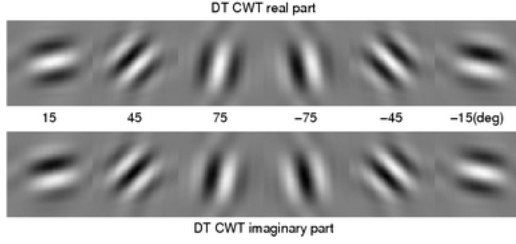


Fig. 1. The real and imaginary impulse responses of the DTCWT. The DTCWT provides 6 directionally selective filters.

provided by the DTCWT is useful to design a keypoint energy measure that emphasises the presence of a keypoint (blob, corner, junction) while ignoring edges and uniform areas. This is an advantage over the Difference of Gaussian detector which requires an extra step to suppress edge responses. The following keypoint energy measure that we propose is based on the product of all six subbands magnitudes:

$$E(C) = \alpha^s \left(\prod_{b=1}^6 \rho_b \right)^\beta \quad (1)$$

Parameters α and β control the relative weight of scales in the accumulated map (see below). Setting low values for α and β will emphasise fine scales and improve the localisation and detection of fine scale features, but will make the detector more sensitive to noise. In our experiments, we found $\alpha = 1$ and $\beta = 1/4$ give the best results on different types of images. We produce m decimated energy maps M_1, \dots, M_m by calculating $E(C)$ for all the coefficients at each scale of the DTCWT decomposition. The number of scales is determined so that the coarsest map is at least 7×7 (e.g. $m = 5$ for a 256×256 image).

3. LOCALISING KEYPOINTS IN THE ACCUMULATED ENERGY MAP

We created a test image (figure 2, left) which comprises various salient features (corners, blobs and a square) of different sizes. Note that a small blob is nested within the large blob, in the lower right part of the image. Since its content is basic, it is easy to judge the relevance of the detection based on our perception.

The m energy maps $\{M_s\}$ previously obtained are decimated by respective factors 2^s . If we detect maxima in these decimated maps (as in [3] with the DWT keypoints), we will obtain keypoints which are poorly localised in the original image space. In SIFT, each local maximum is interpolated using its neighbour values by fitting a quadratic surface and considering its peak location.

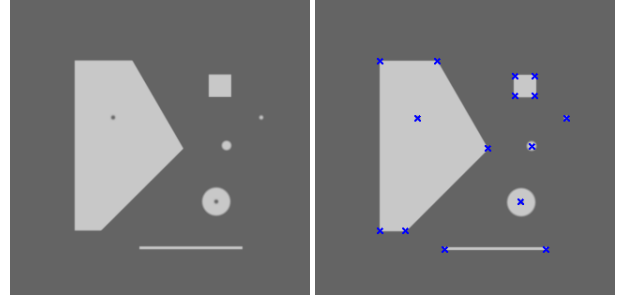


Fig. 2. Left: the input test image with features of different size and nature. Right: the 15 maxima detected from the accumulated map.

We introduce a rather different method to obtain accurate keypoint localisation from the decimated maps $\{M_s\}$. Given a map M_s , we denote as $f_s(M_s)$ the 2D gaussian kernel interpolation up to the original image size (i.e. upsampling by a factor of 2^s). In an interpolated map $f_s(M_s)$, a high energy keypoint produces a high gaussian peak whose variance and inverse curvature are proportional to the scale factor 2^s squared. We define the *accumulated energy map* of the image as the sum of the interpolated maps from scales 1 to m :

$$A = \sum_{s=1}^m f_s(M_s). \quad (2)$$

As a result, the accumulated map is a mixture of gaussians centered about each saliency. The accumulated map corresponding to the figure 2 test image is shown in figure 3 as a surface plot.

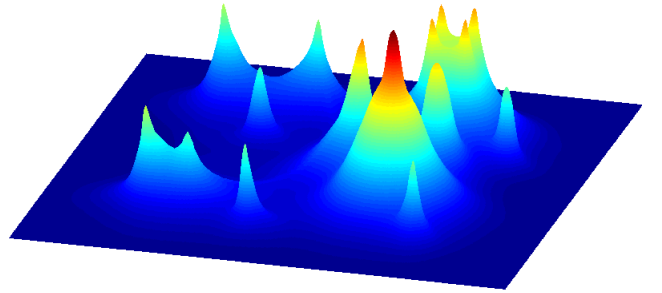


Fig. 3. The accumulated map shown as a surface plot, obtained from energy maps from scales 1 to 5. Its regularity guarantees robust keypoint detection. A high peak indicates that a feature is very salient and/or present at multiple scales. A wide peak indicates the presence of a coarse feature.

Now we define the keypoint locations as the peak locations in A , by simply detecting where energy values in A are maximal on a 3×3 neighbourhood. Thanks to the gaussian interpolation in the construction of A , keypoints are accurately detected at the original pixel resolution. Figure 2 shows the 15 detected maxima on the test image. All the detected locations

do correspond to the perceived saliencies of the image content. This simple maximum detection mechanism is fast but sensitive to noise. To reduce the number of false maxima, we denoise the DTCWT coefficients by the Bivariate Shrinkage Denoising technique [5], before computing keypoint energies. This operation results in a smoother accumulated map.

4. ROBUST SCALE SELECTION

Determining the precise scale of a keypoint is important to define the support region from which an appearance-based descriptor can be extracted for higher level application.

The challenge we address here is the robust selection of keypoint scales from the few scales given by the dyadic decomposition. By comparison, SIFT requires three times as many scales. Interpolating energies across a few scales (typically 5 or 6) and picking the maxima to define the scale of a keypoint is not robust with the dyadic decomposition.

Instead, we exploit the regularity of the accumulated map A and define the scale of a keypoint from the minima of the gradient of A in its vicinity.

Let $g(x, y) \in \mathbb{R}^2$ denote the gradient vector of A at (x, y) . For a keypoint k , we consider the eight projections $p_{i=1,\dots,8}$ of $g(x, y)$ along the following directions around k : $0^\circ, 45^\circ, \dots, 315^\circ$ up to a distance R from k (the radius R is set to 30 pixels here). As a result, given $0 \leq j \leq R$, $p_i(j)$ gives the gradient value along direction i at j pixels away from k . Since a keypoint is located at a peak of A , $p_i(j)$ will be negative for the low values of j and become positive if another keypoint exists in its vicinity. To ignore the interaction from neighbour keypoints, we truncate the gradient projections by setting them to zero from the point they become positive. Note that an isotropic keypoint, such as a blob, will yield similar projections with a strong minimum. A strong feature will yield a strong minimum. Figure 4 shows the eight projections corresponding to the two nested blobs (at the bottom right of the test image). The presence of two blobs (a small and a big one) yields two lobes. The key idea is to define the keypoint scales as the loci where most projections have minima. In presence of noise or complex keypoints, the projections may differ and we need a robust scheme to detect minima which are consistent with all eight projections.

We fuse the eight projections by simply taking their sum and refer to it as the *gradient profile*. The robust detection of its minima is achieved by a flood-filling operation between its maxima: the keypoint scales are defined as the loci that equally split the filled area into two equal areas (see figure 4). The filled area indicates the strength of the keypoint and areas below a threshold τ_{gp} are ignored. In figure 4, the two valid scales are indicated by the dashed lines. The small blob and the big one are assigned scales 4.4 and 19.8, respectively. This scheme naturally allows us to detect the presence of two or more keypoints at the same location which have different scales.

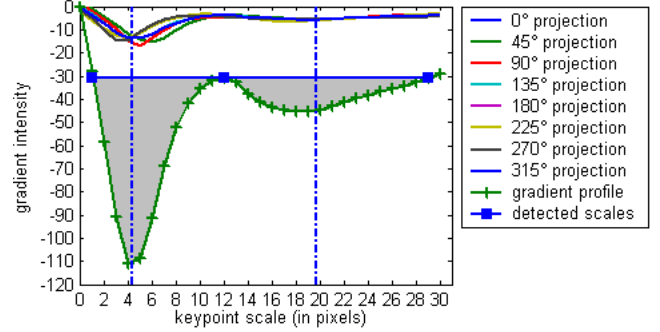


Fig. 4. Scale detection based on the gradient profile and projections determined around the center location of the two nested blobs. Two keypoints of different scales (4.4 and 19.8) are detected here.

5. RESULTS

The performance of our technique was compared against the recent and widely used DoG-based SIFT detector [2].

Perceptual observations: Figure 5 (left) shows the 16 final detected points for the test image as blue circles, from the 16 detected locations (see figure 2). The circle radius indicates the keypoint scale. We observe that all detected keypoints are coherent in scale and position. The scale selection scheme created one extra keypoint for the nested blobs. For

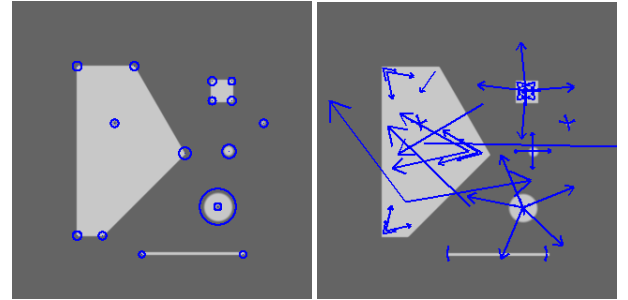


Fig. 5. Detected keypoints with our method (left) and SIFT (right). Circle sizes indicate the keypoint scales. With 16 detected keypoints our method picks the various features, while the 71 detected SIFT keypoints miss some corners.

comparison, SIFT descriptors are shown in figure 5 (right). The arrow lengths indicate the keypoint scale. It detects 71 keypoints total. Figure 6 shows another example of detected keypoints on the standard “cameraman” picture. More generally on different images, we observed that SIFT produces significantly more keypoints and many often refer to the same feature. It often fails to detect coarse corners. Our detector usually detects one keypoint per salient feature, which avoids producing too many redundant keypoints. This is a desirable property as keypoint matching techniques tend to be computationally demanding.

Robustness to rotation: To assess the robustness of our

keypoint detector against SIFT, we ran both algorithms on rotated versions of the test image from 1 to 360 degrees. SIFT detected keypoints on the blobs with great accuracy and robustness but much less on corners. Our method proved to be very robust and accurate for all keypoints at all orientations. To quantify this robustness, we measured the overlap between all detected keypoints at orientation 360° and keypoints detected at all orientations θ° after “derotating” their coordinates by $-\theta^\circ$. The overlap between two keypoints is given by the ratio of twice intersection keypoints discs (whose radius is the scale) by the union of the discs. The total overlap between two images is normalised by the number of keypoint pairs which had an overlap greater than 0.5. Figure 7 shows the plot of this measure for both detectors. This measure remains closer to one across orientations for our detector, indicating a better robustness.

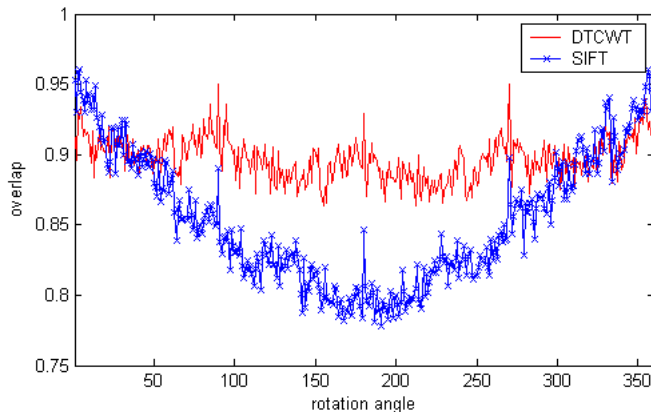


Fig. 7. Repeatability against rotation. The keypoint overlap for our detector remains closer to 1 across orientations than SIFT, which indicates a better robustness to rotation.



Fig. 6. 72 keypoints detected on the “cameraman” picture.

On the contrary, as shown in figure 8, SIFT produces more keypoints as the noise factor increases.

Speed: Our (unoptimised) code runs in Matlab on a 3GHz PC. On a natural 512×512 image, the multiscale keypoint detection takes on average 7s. For comparison, the SIFT [2]

Robustness to noise:

Noise is a critical issue for general keypoint detection since a noisy pixel can be easily interpreted as a fine salient feature. Our denoising scheme (section 3) reduces the fine scale energies to provide a smoother accumulated map. As a result, in presence of strong noise, only coarsest features will be detected. On the contrary, as shown in figure 8, SIFT produces more keypoints as the noise factor increases.

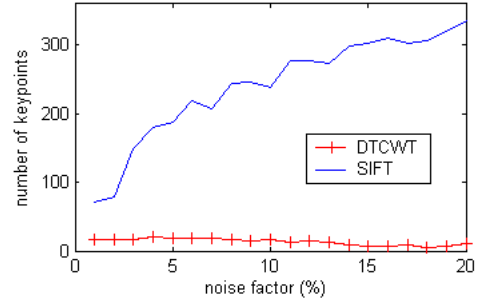


Fig. 8. Number of keypoints detected in presence of noise. In presence of strong noise, our algorithm detects only coarse features while SIFT detects more keypoints.

program (C++ code, version 4¹) takes 4.5s for detection and description of keypoints.

6. CONCLUSION

We presented a novel method to perform robust multiscale keypoint detection based on the DTCWT transform. Both localisation and scale selection operations are based on the accumulated map of keypoint energies. Local saliencies of an image are detected and their scale gives their support region. In comparison with the SIFT detector, results showed our keypoint localisation is more robust to rotation. We obtain about the same order of magnitude in speed. Detected keypoints are less numerous and less redundant (one feature yields no more than one keypoint). Our first results show that the detected keypoints are perceptually consistent with the visual content of the image.

Our future work will focus on the design of multiscale descriptors for these keypoints using the DTCWT.

7. REFERENCES

- [1] C. Harris and M. Stephens, “A combined corner and edge detector,” *Alvey Vision Conference*, pp. 147–151, 1988.
- [2] D.G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [3] E. Loup, N. Sebe, S. Bres, and J-M. Jolion, “Wavelet-based salient points for image retrieval,” *IEEE International Conference on Image Processing*, 2000.
- [4] 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, May 2001.
- [5] L. Sendur and I. Selesnick, “Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency,” *IEEE Trans Sig Proc*, vol. 50, no. 11, pp. 2744–2756, 2002.

¹<http://www.cs.ubc.ca/spider/lowe/keypoints/siftDemoV4.zip>