Convergence of methods for Multi-Scale Image analysis for Object and Image recollection

F. Ferreira J. Dias * V. Santos
Department of Mechanical Engineering * *Institute of Systems and Robotics*
University of Aveiro, Portugal *University of Coimbra, Portugal*

Abstract. There has been a convergence in the implementation of the scale-space methods and schemes that utilize pyramids to store and process image data. In this article we present a fusion of the coherent theory for multi-scale analysis using scale-space and the attractive results in image comparison and object/scene recollection obtained using pyramids. Still in its initial stages, with implementation in a prototype environment partially complete, good preliminary results demonstrate the possible inclusion of such a scheme in robot navigation. Preliminary results from the inclusion of depth data in the process of extraction of interesting points in the image have also been presented.

1 Introduction

The last fifteen years or so have seen the development of methods that treat information in images at multiple scales. Rather than an extension of the noise-versus-information principle behind edge detection methods [1], linear scale-space methods, [2], [3], [4], [5], push a simultaneous multi-scale representation of an image through its embedding in family of functions with a single, continuous parameter.

The difficulty of problems in vision relating to image interpretation, matching and recognition are well known. It lies in the ability of reliably extracting and describing interest points (for subsequent scene or object recognition) in an image taken under variable viewpoint and illumination conditions. Whatever the application, we are drawn to the question of what image properties are invariants. In Schmid and Mohr [6], it is mentioned that the only image invariants are the curvature of the isophote line and the flow line. This same article goes on to suggest that this property is of little practical value given that the associated calculations are difficult, and that the noise in real images and their limited resolution negates the remaining usefulness of the result.

Our interest in multi-scale image analysis has to do with a robot mission programming approach currently being developed at our mobile robot laboratory that uses object recognition to trigger consecutive mission phases.

2 Linear Scale-space

Linear scale-space has been often presented as a good analysis tool for early-vision with parallels being drawn with results from research in animal vision. The goal of multi-scale analysis of 2-D images is to progressively blur(smooth) out fine details so that other, more subtle, image features begin to dominate. This blurring is carried out by embedding the image

in a family of functions related through the variation of a single parameter. It is important that the blurring process must never pollute the image with spurious information.

A third axis, the scale parameter, is grown from the image coordinates, the "scale-axis", which allows the possibility of ordering features in scale and performing measurements along this dimension. The aim of the approach becomes to 'optimize' the detection of features of interest in the image across scales. A larger value of the scale refers to a greater smoothing of the image.

The embedded function, $L_{x^i y^j}$ of the image intensity is obtained by convolving it with a kernel. Important considerations for the smoothing kernel used include

Causality: Any structure that is present at a larger scale must have existed at a smaller scale. The application of the smoothing kernel must lead to the progressive destruction of features without the addition of spurious detail at any level.

Ability to apply smoothing in a cascaded form: In order to treat every scale in a similar manner and in the interests of computational efficiency, it must be possible to reach any value of scale from any other, equation(1). The behavior of the same features in amplified image must be related in some way to the scale.

$$L_{x^i y^j}(.,.,t_2) = G(.,.,t_2 - t_1) * L_{x^i y^j}(.,.,t_1) \tag{1}$$

Homogeneity and Isotropy: Every pixel must have the same potential of affecting the behavior of the image in scale space. This translates into the requirements for an odd-size kernel subject to other limitations such as the existence of a single mode and symmetric values about this mode. The idea of an uncommitted front-end means that no preference must be given to features oriented in one direction. Invariance to rotation is affected by the degree of isotropy of the kernel and of the interest point detectors.

The Gaussian kernel has been shown to be the unique kernel that satisfies the criteria outlined earlier. If we choose the maxima of the illumination function embedded, in the Gaussian kernel, along scale as the feature of interest, the causality criterion leads us to the conclusion that, at least a part of, the function must lie at lower scales than the one at which the maxima was observed. The diffusion equation (2) allows the behavior of the derivative functions of $L_{x^i y^j}$ in image coordinates to be used as proxy for the behavior of $L_{x^i y^j}$ in scale space,t. The same relation is found to hold also for functions of $L_{x^i y^j}$, equation (3). Another important property of the intensity image that is now embedded in the Gaussian kernel is that is always infinitely differentiable

$$\frac{\partial L}{\partial t} = \frac{1}{2}\nabla^2 L \tag{2}$$

$$\frac{\partial \mathcal{D}_{x^i y^j}(.,.,t)}{\partial t} = \frac{1}{2}\nabla^2 \mathcal{D}_{x^i y^j}(.,.,t) \tag{3}$$

Lindeberg utilizes polynomial functions, homogeneous in order of derivatives, to identify image characteristics. The conversion of the single image coordinate system to local gauge coordinates allows much flexibility in the comparison of the derivatives and functions (N-Jet) of these same derivatives. The approach consists of an automatic Scale-selection, that seeks to match features with particular scale levels by selecting the scale at which the features assume maximum values along the scale dimension. Through the diffusion equation 3, maxima along scale are detected through maxima in space. Normalization of the N-Jet across scales enables

a comparison between values of the smoothed signal itself and consequently the derivatives *across* scales.

Articles by Lindeberg such as [7], [8], [9] deal with the development of a parallel theory that seeks to develop extensions of the scale-space theory for the discrete image case. Through an extension of a 3-point(in 1-D) smoothing another function containing the modified Bessel functions of integer order, equation (4), is developed as the 'discrete analog of the Gaussian' for increasing scale, t.

$$T(n,t) = e^{-t} * I_n(t) \tag{4}$$

The diffusion equation also leads to the relation such that the detection scales for the same object in a zoomed image f' is related to detection scale in the original image f through equation (5), s being the zoom applied.

$$f(x,t) = f'(sx, s^2t) \tag{5}$$

3 Invariant point descriptors in Multi-scale image analysis

The SIFT(Scale-Invariant Feature Transforms) method, developed by Lowe, creates histogram -based descriptors that attempt to be invariant to changes in the illumination, viewpoint, translations and partial image occlusions. The gradient of Difference of Gaussian(DoG) images are used to select interesting points. Lowe [10], has utilized the method for a host of applications that use matching to achieve various ends. These maxima features are located by searching for the maxima of the DoG in space and pyramid depth followed by a filtering procedure to limit the response to edges and points of low contrast. Successful applications range from limited 3-D localization, stereo matching of points to challenging object identification in a scene.

Polling of local gradient magnitudes and directions at points around the interest point create a large discerning vector description that is invariant to rotation and translation in the image plane. Large tolerances for feature properties at the time of matching are acceptable due to the high dimensionality of this description vector. Coarse, low frequency features are favored over higher finer features, representing a protection against noise and small variations in image.

Lowe [11] describes algorithms for data organization and the fusion of features from different viewpoints that expand the object recognition capabilities (3-D rotation) of the method. Through the tracking of multiple views of the same object, larger invariance stability of object and scene detection is obtained. Decision theory and conditional probability are also involved [12]. In another application, Brown and Lowe [13] utilize the SIFT method to perform a panoramic arrangement of images that have some overlap.

Schmid [6] constructs a "Local-jet" at every point in the image, the points of interest being chosen according to a Heitger and Rosenthaler detector. In Schmid and Mohr [14], the same authors utilize a Harris detector to choose interest points in an image. Mikolaczyk and Schmid [15], utilize other functions to extract points in an image and perform selection along the scale dimension. They found the Harris function $(det(C) - \alpha \text{trace}^2(C))$ to be the most insensitive to image rotations, illumination perspective changes.

Hall [16]use a YUV space to create a vector based on higher order luminance terms to distinguish structure and another based on chrominance to resolve ambiguity. The interest point detector selects points based on the ambiguity of the vectors and depend on the content of the entire scene and training set, rather than simply choosing local maxima.

In Mikolajczyk and Schmid [17], a comparison is made between various methods of extracting points of interest and the efficacy of a few implementations of invariant region descriptors. Comparisons were made various with descriptors including the affine invariant descriptor developed by them, steerable filters, cross correlations and differential invariant and the SIFT feature descriptor developed by Lowe. Only in the face of illumination changes did the SIFT operator performance not achieve the best rating. It is common to see point descriptors based on histograms that have been 'normalised' in orientation to provide invariance to rotation.

4 Hybrid-Scale-Space and pyramids

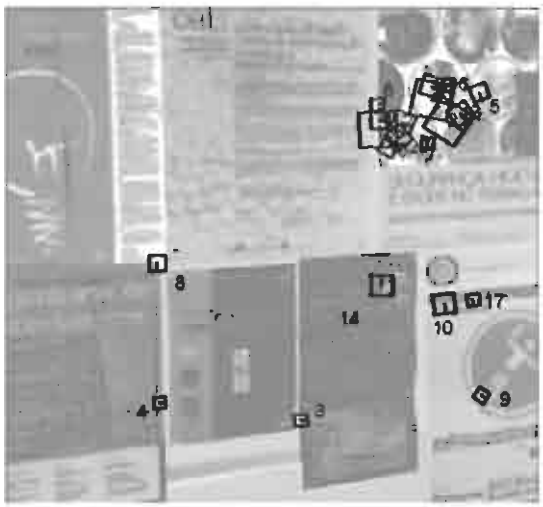
Pyramids are multi-resolution data structures that have been utilized for quite some time in order to take advantage of memory savings and increased processing time as information redundancy in an image increases. Image pyramids to store image data that has been consecutively smoothed by a Gaussian function have been used by Burt [18], Crowely [19] and Lowe [10], among others.

In [20], Lowe introduces approximations to the scale-space theory through the use of sub-sampled pyramids(a reduction in resolution does not follow every smoothing step). Normalization is addressed as being necessary in order to compare extrema across scales (though it is circumvented through a suitable choice of scale increments). The number of smoothing and feature search operations per octave has been arrived at, empirically, as '3'. A localization step is introduced after the feature detection stage. The efficacy of a large feature descriptor for the storage of a great deal of information is demonstrated through the implementation of a 128 dimensional K-d tree for matching a database with hundreds of thousands of points. Feature extraction is followed by the matching in this K-d tree and confirmation using a variant of the Hough transform followed by a procedure outlined in [21] for transformation parameter estimation.

In recent work, Lindeberg [22] utilizes sub-sampled pyramids as a data reduction method that enables use of Linear scale-space in real-time applications. Introducing the generalized "Hybrid Scale-Space", the article describes a multi-scale image analysis method that can be applied both in the case where resolution is maintained or where it is reduced as in the case of pyramid representations. The application of γ - normalized derivatives polynomials to extract image features at the their 'optimum' scale is presented for this general case with construction details. The focus remains on the isolation of image features such as edges and ridges within the image.

In this work we attempt to use the image matching method strategy developed by Lowe. It consists of 1) creating the local N-Jet function at various scales with decimation 2) feature extraction, 3) feature description through the construction of a large vector description 4) Matching using a modified K-d tree. We have used the discrete analogue of the Gaussian kernel developed by Lindeberg to create the multi-scale representation. Since the convolution commutes with the derivative operator, the intensity image has been convolved followed by the calculation of the derivatives.

We shall be using an N-Localjet function in order to characterize our points of interest in terms of specific geometric features. This, we expect, will allow us to have better control on the type of features we select. While we have used 'Junction detectors', $L_y L_y L_{xx} - 2L_x L_y L_{xy} + L_x L_x L_{yy}$, developed quite thoroughly by Lindeberg in [23] in the present work, we are working to develop stable detectors for other types of image features. Experiments were conducted with a 5-bin-3 and a 5-bin-5 pyramid(3 and 5 levels of smoothing before a reduction of



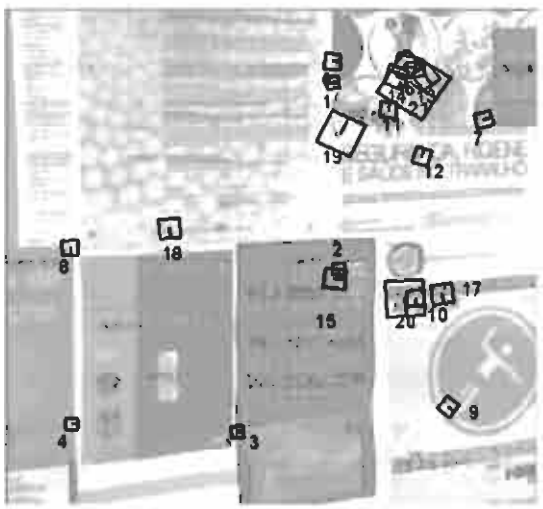
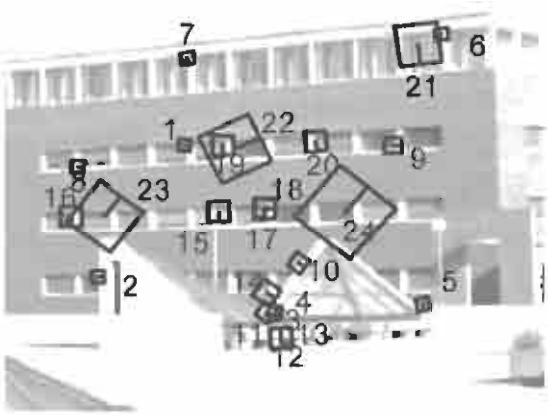


Figure 1: Interior scene taken in artificial lighting conditions.



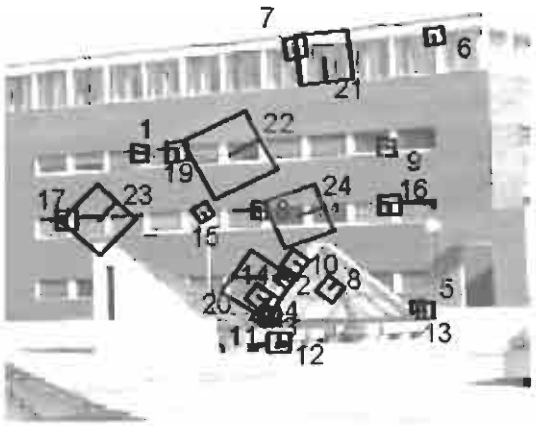


Figure 2: Exterior scene with a building having number of similar architectureal features.

resolution).

A combing of scales, as described in [23] was carried out after feature detection to obtain the scale at which the feature can be best represented (less than or equal to the scale of detection). This localization procedure balances the increased noise present at finer scales with the effects of bad positioning due to feature interference inherent at coarser scales. The 128-long description vector similar to the SIFT descriptor was created. At this stage we have attempted to match only the 100 best points(strongest Junction response) of pairs of images.

While the search time of Kd-trees cannot be decreased from $O(\sqrt{n}-k)$ time [24], heuristic methods can be found to speed up the search by constant amounts. This has been done to some extent by Lowe, as described in the matching section in [20].In this present work a 128 element long vector that describes the histogram of gradient orientations at 16 neighboring points of the point of interest is utilized. Each of these 128 values are given equal weights and a K-d tree with up-to 128 different dimensions is built to accommodate the database of interest features from all the models.

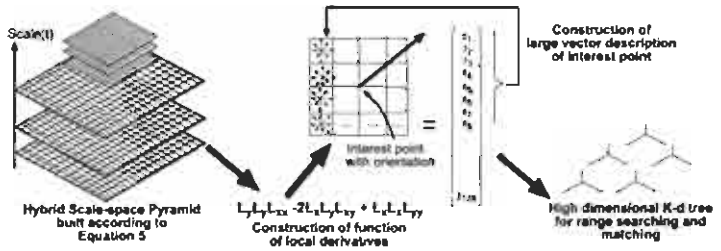


Figure 3: Flowchart: Pyramid construction and point extraction follow from linear scale space; Feature point description and matching are due to Lowe.

5 Experiments

Experiments were performed on two sets of images, figures(1) and (2). These, represented in the first row, consist of one taken of an indoor scene with artificial white light and a second, in bright sunlight of a building containing a number of very similar architectural features. The algorithm for extraction of points and the creation of a vector description was run on images represented in the second and third rows, the latter being an amplified image. The pictures(640x480pixels) are taken with a hand-held camera and amplification has been achieved optically. The images are, therefore, subject to some general transformation, rather than a simple scaling and translation.

In figure (1) some 'matched' interest points in the right image have not been located at positions corresponding to their position in the right image. As an example, feature marked 14 in the second row image has been matched with a feature in another region, that clearly is incorrect.

In figure(2) the matching results are affected by the presence of multiple similar geometry features in the image resulting in some ill-matching. Since no location constraints has been applied in the matching procedure, some features (feature marked 1, 8, 16 and 24 for example) have been matched with other features arising out of similar geometry of regions in the image. This suggests that the matching process must include some spatial distribution constraints on the matching that points can achieve.

In both sets of images we notice the scale-space behavior of the points in which the corresponding in the zoomed image are detected at higher scales, equation (5).

6 Conclusion and Future work

We expect improvements in the results of the application of the method to come from modifications to the vector description of the feature points and the procedure utilized in matching these descriptions. The implementation does not distinguish between differences in the gradient orientations around the same point and gradient orientations between different points. Reducing the effect that within-point gradient distribution differences might have, compared to differences in the gradient distribution at different points might create a better discerning vector.

Getting the 128-vector description to distinguish between a large group of points depends on whether the 128 values do, actually, take very distinct values. If the large discerning power of this 128 term vector is not utilized in practice, the large length of the vector would only be a liability. In particular, performing normalization during the creation of the vector destroys the orthogonality of those dimensions which would create problems during matching. Some other procedure must be found that creates the same effect as normalization, i.e. invariance to illumination and noise, but maintains the independence of the dimensions of the search tree.

Besides the improvements in the range-searching procedure, other feature detectors are being developed with an aim to achieve object and scene description through multiple feature extraction. These geometricaly interpretable features shall be described through N-Jet functions similar to the ones developed by Lindeberg with an emphasis on features that allow the creation of descriptions that can be matched across transformations in 3-D.

References

[1] V. Torre and T. A. Poggio. On edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:147–163, 1986.

[2] Jan J Koenderink. The structure of images. *Biological Cybernetics*, 50(5), 1984.

[3] A.P.Witkin. Scale-space filtering. In *Proc. 8th Joint conference on Artificial Intelligence*, pages 1019–1023, Karlsruhe, W. Germany, 1983. 8th Joint conference on Artificial Intelligence.

[4] Tony Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Press, 1994.

[5] Bart M ter. Haar Romeny, editor. *Geometry-Driven Diffusion in Computer Vision*. Kluwer Academic Press, 1994.

[6] Cordelia Schmid and Roger Mohr. Matching by local invariants. Technical report, INRIA, 1995.

[7] Tony Lindeberg. Scale-space for discrete signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-12(3):234–254, 1990.

[8] Tony Lindeberg. Discrete derivative approximations with scale-space properties: A basis for low-level feature extraction. *Journal of Mathematical Imaging and Vision, JMIV*, 3(3):349–376, 1993.

[9] Tony Lindeberg. Edge detection and ridge detection with automatic scale selection. Technical report, 1996.

[10] David G. Lowe. Object recognition from local scale-invariant features. In *Proc. of the International Conference on Computer Vision, Corfu*, pages 1150–1157, 1999.

[11] David G. Lowe. Local feature view clustering for 3d object recognition. Kauai, Hawaii, 2001.

[12] Arthur Pope and David G. Lowe. Probabilistic models of appearance for 3-d object recognition. *IJCV*, (40,2), 2000.

[13] Brown M. and Lowe G. David. Recognising panoramas. In *Tenth International Conference on Computer Vision (ICCV 2003)*, October 2003.

[14] Cordelia Schmid and Roger Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997.

[15] Krystian Mikolajczyk and Cordelia Schmid. Indexing based on scale invariant interest points. In *International Conference on Computer Vision*, Vancouver, 2001.

[16] Daniela Hall, Vincent Colin de Verdiere, and James L. Crowley. Object recognition using coloured receptive fields. In *ECCV (1)*, pages 164–177, 2000.

[17] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. In *CVPR03*, pages II: 257–263, 2003.

[18] Burt P. Fast filter transforms for image processing. *Computer Vision, Graphics and Image processing*, 1981.

[19] James L Crowley and Richard M. Stern. Fast computation of the difference of low-pass transform. Technical Report CMU-RI-TR-82-18, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, November 1982.

[20] David G. Lowe. Distinctive image features-from scale-invariant keypoints. *International Journal of Computer Vision*, 2003.

[21] M. Brown and Lowe D. Invariant features from interest point groups. In *British Machine Vision Conference, BMVC 2002*, Cardiff, 2002.

[22] Tony Lindeberg and Lars Bretzner. Real-time scale selection in hybrid multi-scale representations. Technical report, Department of Numerical Analysis and Computer Science, KTH, 2003.

[23] Tony Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):77–116, 1998.

[24] M. de Berg, M. van Krevald, M. Overmars, and O. Schwarzkopf. *Computational Geometry, Algorithms and Applications*. Springer Verlag, 1997.