Tina Memo No. 2009-003
Submitted to BMVC 2010 (rejected)
Presented at the Experimental Psychology Society meeting, Bangor July, 2013, see also memo 2010-003.

Dual Component Linear Interpolation of View Spheres.

N. A. Thacker, S.Coupe and P.A.Bromiley

Last updated
8 / 05 / 2010

Imaging Science and Biomedical Engineering Division,
Medical School, University of Manchester,
Stopford Building, Oxford Road,

Manchester, M13 9PT.

Dual Component Linear Interpolation of View Spheres

Abstract

This paper describes an approach for the representation of projected 3D edge features for purposes of view-based recognition and localisation of objects. It is based upon the representation of arbitrary configurations of features using geometric co-occurrence as Pairwise Geometric Histograms (PGH). Sets of these histograms can be shown to provide a complete representation for arbitrary configurations of fixed 2D projections of 3D shape. Topological considerations allow us to determine the local variation of these descriptors as a function of viewpoint and scale. We describe a mathematical model for the interpolation of correlated changes in these histograms, comprising two independent linear models for use during simultaneous view and scale matching. Mismatch and match distributions are provided to give a context for the accuracy of approximation. An assessment is made of the utility of this approach for object localisation (more results are presented in a companion paper).

Introduction

Early work in computer vision [11] suggested that it would be infeasible to recognise 3D objects without using view-based invariant features. Marr argued for the extraction of 3D data as the ideal approach to construction of an invariant shape representation. However, the motivation for attempting a view- based approach was bolstered by Rosenfeld [3] and a number of studies which indicated that humans only store specific views of objects for the purposes of recognition [4, 15]. It was suggested that this could be mediated by a representation of local geometry [8].

It has been found possible to construct local view-based representations with a number of invariance characteristics without necessarily extracting a 3D representation. Currently popular approaches such as SIFT [5] go some way to achieving these ends by computing intermediate representations based around focus points that are partly invariant to orientation/scale and illumination. More recently, efforts have been made to evaluate a growing number of similar methods [10]. Such techniques are useful for generating rapid clues as to image content. However, many objects do not provide many (if any) feature points, making this approach unsuitable. It can be argued that a solution to the problem of general purpose recognition will require the ability to encode shape. Historically, edges have always been seen as the most suitable way to represent the fundamental structure visible in projected views of objects. Consequently, attempts have been made to recognise (or index) objects based upon edge contours [9, 13]. Limiting the possible combinatorial problems associated with variations in projected shape requires careful consideration of the role of invariances. This must be done while also considering the effects of occlusion on any chosen representation, if the resulting methods are to have any real utility for arbitrary scenes. This raises the question; could there ever be an optimal representation that encodes some maximum number of possible invariances in a statistically meaningful way?

This question was at least partly answered in the 1990's by research that listed the possible invariances and tried to select a subset that would support unambiguous recognition of edge-based boundary shape based upon statistical principles. The conclusion was that an "in plane" rotation and shift invariant representation, with some degree of illumination invariance, could be constructed from which the original object shape could be recovered (i.e. completeness [12]). It can be concluded that issues such as scale and "out of plane" 3D rotation cannot be directly addressed due to the varying statistical stability of computed invariant quantities. Such instabilities would destabilise subsequent recognition and also destroy representational completeness; a property essential for recognition of arbitrary objects without representational ambiguity. This work stopped short of recognising 3D projections of objects, due to the massive storage and processing requirements (correctly predicted by Marr), in comparison to the processing capabilities available at that time. In the years since then, the capabilities of commercially available hardware have continued to increase exponentially in accordance with Moore's Law. This now offers at least a possibility for view-based recognition of projected 3D shape using a complete representation. It is the aim of this paper to investigate the feasibility of this idea.

The method of mixture modelling is often used for the description of extended distributions. Generally, these representations are based upon the summation of Gaussian or other localised distributions. Such

approaches work well as part of an appearance model approach to object localisation. However, we argue here that such distribution models are topologically inappropriate for view-based recognition. We describe an alternative that better conforms to the expected behaviour and test this model quantitatively.

Methods

Provided that care is taken to construct data representations that change smoothly as a function of the image formation parameters, visual data must always generate representations that lie on low dimensional manifolds (i.e. a 2D view sphere, of a rigid 3D object). These can be characterised as being non-linear (due to projective geometry), locally continuous but globally discontinuous (due to occlusion) and self intersecting (due to symmetry) [2]. Although mixture models have become popular as tools for solving general pattern recognition problems, the topological behaviour of view spheres can be better described as an extended lamina manifold. The effects of scale produce additional variation which itself is part of a continuous non-linear 1D variation. Given the extended nature of both sources of variation, it seems to us inappropriate to model all of these effects as a localised distribution. We suggest instead a mechanism for representing a group of node points as a series of connected hyper-surface approximations, followed by a second (non-orthogonal) density model incorporating the effects of additional variations such as scale. We require a general purpose non-linear classification system that can describe this manifold in order to build a shape recognition system for rigid 3D objects. This is the starting point for what follows.

The properties listed above, combined with appreciation of statistical constraints, tell us that in general we would not expect to be able to build a global linear model of a view sphere. More sophisticated density models, such as the popular non-linear representations (e.g. kernel PCA), simplify the modelling of specific decision boundaries, but in doing so distort the statistical character of representation spaces (by non-linearly propagating the effects of noise). In order to avoid this we must stay with localised linear approximations of the manifold (Figure 1). Piecewise linearisation therefore has theoretical advantages, as it supports a simpler approach to the statistical estimation of parameters. The intention here is to construct a computationally efficient representation scheme that can be logically justified as a model of view variability and utilises a generically valid similarity measure.

For recognition, we work with a histogram-based representation of local geometry. This is based upon density estimation of local geometric co-occurrence of relative orientation and perpendicular distance, in accordance with [12], or Pairwise Geometric Histograms. The task of recognition therefore becomes that of matching histograms.

Recent publications have suggested that for cases where the data generation process induce correlations into the histogram distribution, approaches such as the "earth mover distance" will give better matching performance than conventional methods such as the Chi-square [6, 7]. A novel aspect of our work is that we aim to model these correlations and take direct account of this aspect of variation during matching. With correlated structure removed the remaining component is correctly described by conventional statistics and we use the Bhattacharyya measure [1]. This can be shown to approximate Fisher's exact test for Poisson sampled data, with half the approximation error of the Chi-square statistic (Appendix A).

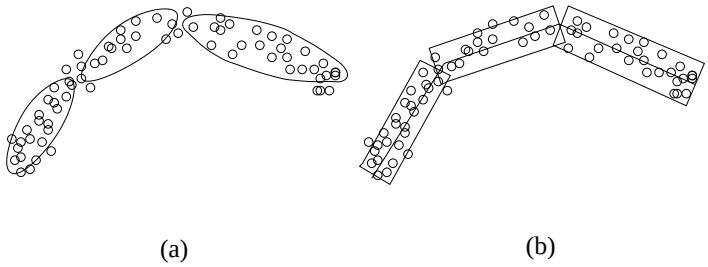


Figure 1: Modelling the view direction manifold with (a) Gaussians and (b) linear sections.

Another novel aspect of this work is that we separate the modelling of view variability induced into geometric histograms into two components, each modelled by separate linear systems (a dual component linear model). This is done in order to allow us to represent variation with a topology which is consistent

with the data generation process (i.e. a smoothly connected view sphere, and smoothly connected scale variation). Although the combination of these two models is a linear model (thereby supporting closed form solutions for similarity comparison), the resulting piecewise approximation across view sphere and scale is not. Also, only the 2D variation due to view dependency is stored (learned) during training. Any other modes of variation are estimated from the incoming data via resampling, thereby reducing the storage requirements (which we believe to be the limiting factor for a large-scale recognition system) at the expense of a more sophisticated matching strategy.

Mathematical Representation

We start with an estimate of the local geometrical co-occurence, as a function of perpendicular distance d and relative orientation θ in the form of a 2D density distribution $p(d, \theta)$ for each line fragment. We then work in the variance normalised space of this data by applying the square-root transform, so that dot products between distributions compute the Bhattacharyya measure. We can approximate the variation of a histogram $h = \sqrt{p(d,\theta)}$ accompanying changing view point by a linearised 2D variation, scaled by weights λ_i i.e.

$$h = \lambda_0 \bar{h} + \sum_{i=1}^{2} \lambda_i e_i$$

where the dimensions of linear variation e_i are orthogonal to the mean direction vector $\bar{h}$. This "tangent space" interpretation arises due to the scale invariance of the histogram matching process, which is in turn needed to account for feature length differences occurring during the linear approximation of object features (i.e. fragmentation). The closest match of this representation to a specific test histogram g is

$$h' = \bar{h}.g\ \bar{h} + \sum_{i}^{2} (e_i.g)\ e_i a \qquad (1)$$

normalised to the unit sphere. The similarity function we will use to match corresponding model parts will be of the form $B^2 = (h.g)^2/|h|^2|g|^2$ in accordance with use of the Bhattacharyya measure (B) for distribution comparison. Therefore, the best match score attainable within the linear model will be

$$\frac{(h'.g)^2}{|h'|^2|g|^2} = \frac{[g.(\bar{h}.g\ \bar{h}) + g.(\sum_i^2 (e_i.g)\ e_i)]^2}{|h'|^2|g|^2} = \frac{|h'|^2}{|g|^2}$$

We now represent the possible variation of the observed data g as an additional linear model. This model too is a tangent space representation, but there is no requirement of orthogonality to the previous components of the system. Using the same approach as above

$$g = \bar{g} + \sum_{k} \delta_k f_k \qquad (2)$$

As the overall system is still in effect a linear model, and the similarity measure is quadratic, we can determine a unique best match for this second model with the view model by determining the δ_i which gives the best similarity score

$$\frac{|h'|^2}{|g|^2} = \frac{[(\bar{g} + \sum_k \delta_k f_k).\bar{h}]^2 + \sum_i [(\bar{g} + \sum_k \delta_k f_k).e_i]^2}{1 + \sum_j \delta_j^2}$$

If we are considering only scale we have a 1D model approximated by the vector δf, we can then write

$$\frac{|h'|^2}{|g|^2} = \frac{[(\bar{g} + \delta f).\bar{h}]^2 + \sum_i^2 [(\bar{g} + \delta f).e_i]^2}{1 + \delta^2} \qquad (3)$$

The value of δ that maximises this expression (i.e. $\partial B^2/\partial\delta = 0$) can be re-written as a quadratic in terms of δ

$$A + B\delta - A\delta^2 + 0\delta^3 = 0$$

with

$$A = \bar{g}.\bar{h} f.\bar{h} + \sum_{i}^{2} \bar{g}.e_i f.e_i \quad , \quad B = (f.\bar{h})^2 + \sum_{i}^{2} (f.e_i)^2 - (\bar{g}.\bar{h})^2 - \sum_{i}^{2} (\bar{g}.e_i)^2$$

and

$$\delta = \frac{B \pm \sqrt{B^2 + 4A^2}}{2A}$$

The vectors h, e_i and f required for these calculations can be easily determined from three histograms computed at the vertices of a triangulated view patch and the scaled versions of the input data. As the above gives two solution candidates we can select the best via direct substitution into equation 3. For this work, we have used a set of 3D wireframe models constructed for 8 randomly selected man made objects, in order to compute ground truth histogram data. Histograms have dimensions of 64/20 corresponding to angles of $0 - 4\pi$ and perpendicular distances of $-50 - 50$ pixels respectively, with commensurate levels of blurring used during construction in order to set the level of shape representation accuracy. We have evaluated the interpolation model quantitatively, considering the view and scale-based variations and their effects on the behaviour of the computed histogram match score in order to evaluate the efficacy of using an interpolation method for view-based representation.

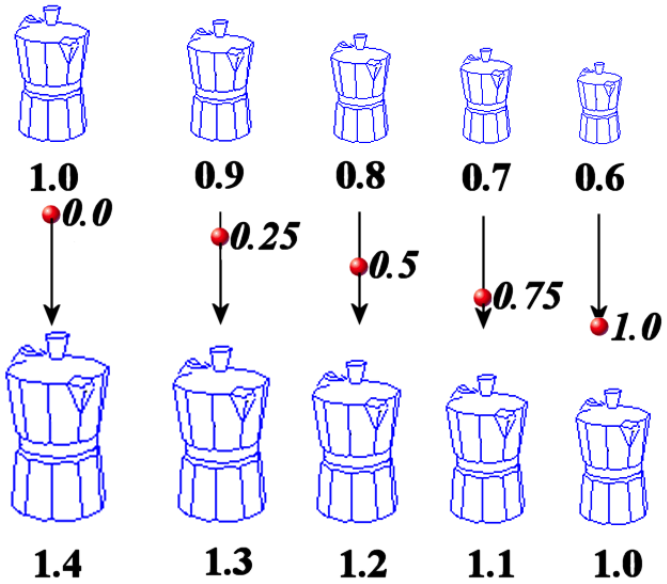


Figure 2: An example of the sets of scaled wireframe data used for the test of scale-based interpolation. For each pair of scales used for construction of the model the scale tested (corresponding to data of one fixed scale) is shown.

Experimental Design

A random view point and reference was chosen for each of 8 objects from which to compute a geometric histogram. Histograms were constructed at the vertices of an (approximately) equilateral triangle centred around this view. In order to eliminate discontinuities due to occlusion, a fixed set of features were selected that were jointly visible at all three locations. An eigen vector analysis of all histograms sampled from within a pentagonal region were used to evaluate the hypothesis that view-based representations of 3D shape should lie on a 2D manifold. This preliminary work indicated that the local variation in histograms varied in the manner expected as a function of view direction (i.e. in a PCA analysis, all but a few % of the variation was found in the first two eigen modes about the mean). The greatest deviation from the linear model always occurred at interpolations close to the centre of the triangle. The difference between Bhattacharyya scores computed for the (ground truth) central view and interpolated from the linear model computed from the vertex histograms were then used as the basis for evaluation of the interpolation strategy over a range of view points.

In a second set of tests, the same set of object wireframes were used to construct data at two scales corresponding to a range of 40% of the original object size (Figure 2). The object scale was estimated

using the theory above to estimate δ from the resulting match scores for the histogram interpolated from the view-based model.

The above analysis concentrates on the approximation of histogram variation using the dual linear model. For such work we could only use simulated data, in order to establish a ground truth for the linearisation processes. In order to put these results into context, we also need to know the match distributions for correct and incorrect matches. To this end, interpolated view spheres were constructed for all objects and matched to real world views containing single objects at scales around 10% of the mean interpolated model. The 24 longest lines were extracted from this scene and their histograms matched to those from the 12 longest line fragments from each of 42 possible views for each object (a compromise on performance implemented in the interest of computational efficiency). Evidence for view directions was accumulated and individual lines contributing to consistent view points were used to initialise a projection of the original object model into the scene. These models were then optimised in order to validate the presence of expected edge features using the method described in [14]. The visually validated matches were then used to identify line features which had contributed to the view hypothesis. Match scores corresponding to these valid lines were then entered into a histogram in order to determine their distribution. The mismatch distribution was constructed separately by matching random objects (of incorrect type) to the image features and entering the match scores into a histogram.

Results

The view interpolation results are presented in Figure 3. They show how the match scores resulting from the interpolation process approximate the theoretical best match value (1.0) as a function of increasing view separation. The interpolation process supports viewpoint modelling over approximately 30 degrees for the level of geometric histogram resolution used, which is at least 3 times more (i.e. 9 times less storage requirement) than using a simple nearest neighbour.

Figure 4 shows the accuracy of scale interpolation and corresponding match scores for a linear interpolation over a 40 % change of scale (see figure 2). Matching 24 features from the scene to the 42 views of an entire view sphere of an individual object generally required several minutes on a ThinkPad with a Centrino2 processor running Linux. Results indicate that the value of δ can be used to infer the scale of an object to within 10 %, and that worst case interpolated match scores are accurate to a similar proportion.

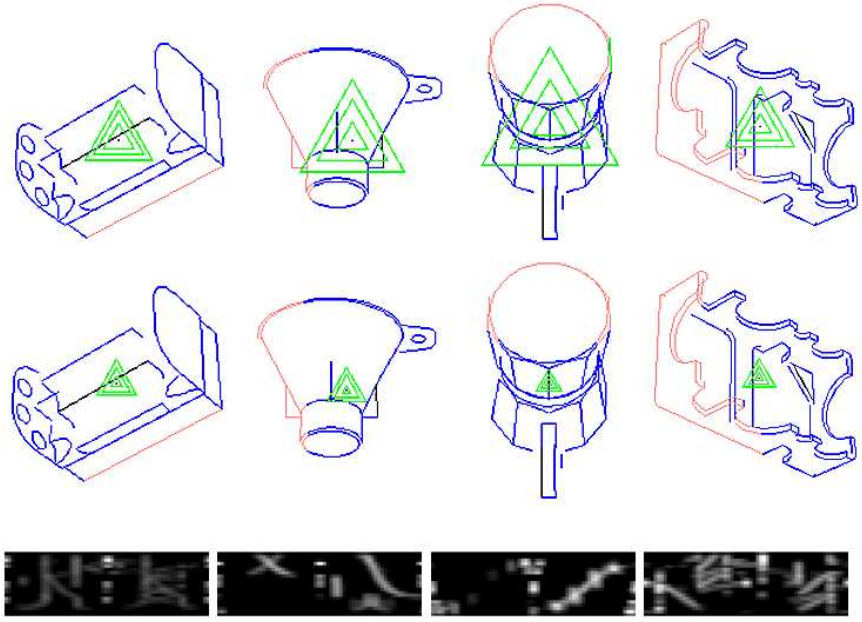
Figure 5 shows the distribution for valid and invalid matches obtained during a view-based recognition experiment. We can see that with the parameters we have selected the correct match distribution is quite well separated from the background mis-matches, allowing individual line hypotheses to be matched quite reliably during the accumulation of evidence for each object view and scale.

Conclusions

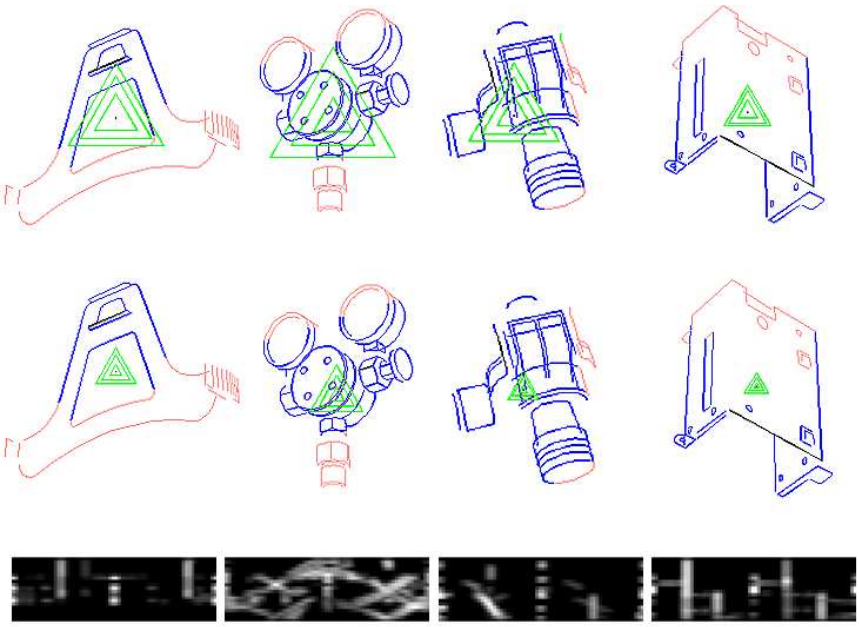
This paper gives a general analysis of the process of modelling projections of objects as a series of views. It explains the rationale for representing local shape as co-occurrence distributions of edge features and a piecewise linearisation of the viewsphere which matches the topology expected from the shape representation. Our results with simulated data demonstrate the accuracy with which the view sphere of a 3D object can be modelled with two linear components for purposes of matching. We have shown that the process is consistent with the requirements of feature-based matching by extracting the distribution of match scores for both correct and incorrect feature match hypotheses in real data. Further quantitative results pertaining to the use of this technique for the localisation of objects using view- based recognition will be presented in a companion paper at this conference.

Appendix A

Although the use of square-root transforms is a long standing technique in statistics, in order to more closely approximate a Poisson distribution with a Gaussian, a proof of its behaviour has been difficult to locate in the published literature. We therefore provide a synopsis of our own derivation here.



(a)



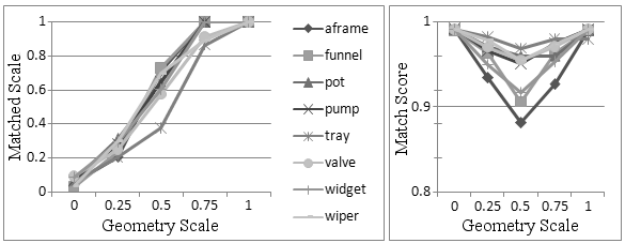
(b)

Figure 3: The objects and their views used in these experiments, along with the corresponding central view histogram for the chosen feature (black). Features in blue show which subset of features in the original object contribute entries to the histogram. In the upper rows, the green triangles show the range of viewpoints corresponding to interpolated match scores of 0.99, 0.95 and 0.9 respectively. The lower rows are the same data without 2D interpolation (i.e. matching to the mean histogram only).

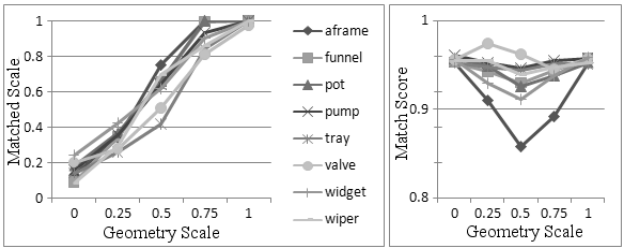
Let $P(x;\lambda)$ be a Poisson distribution with mean λ, and consider the usual Gaussian approximation $G(x|\lambda)$ with $\mu = \lambda$ and standard deviation $\sigma = \sqrt{\lambda}$.

$$G(x|\lambda) = \frac{e^{-(x-\lambda)^2)/2\lambda}}{\sqrt{2\pi\lambda}} e^{h(x,\lambda)} = P(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \tag{1}$$

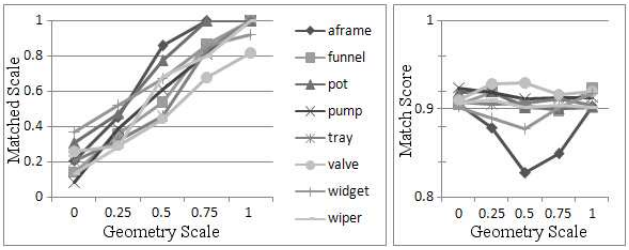
where $e^{h(x,\lambda)}$ is the multiplicative approximation error. This error can be obtained by applying Sterling's



(a)



(b)



(c)

Figure 4: The accuracy of the model for estimation of object scale and the resulting distribution of computed match score for view approximation accuracies of 0.99 (a), 0.95(b) and 0.90(c) over a 40% change of scale (indexed as geometry scale 0-1).

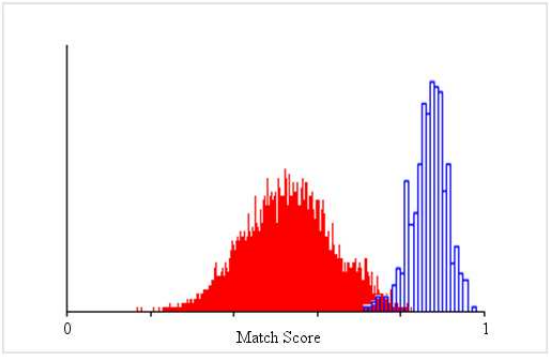


Figure 5: Distribution of match scores from real data for lines validated as being consistent with the projected wire frame model (blue) and all mismatch line hypotheses (red).

approximation $x! = x^x e^{-x}\sqrt{2\pi x}$ to the denominator, taking logarithms of both sides, applying the substitution $\lambda/x = z + 1$, then using the Mercator series expansion for $ln(1 + z)$ and the binomial expansion for $(z + 1)^{-1}$, giving

$$h(x, \lambda) = \frac{(\lambda - x)}{2x} - \frac{(\lambda - x)^2}{4x^2} - \frac{(x - 1)(\lambda - x)^3}{6x^3} + \frac{(2x - 1)(\lambda - x)^4}{8x^4} ... \quad (2)$$

This has large linear factors in $(\lambda - x)/x$ for small x but is dominated by the cubic terms for large x.

However, if the process is repeated for the approximation of the square-root of a Poisson variable (i.e. after changing variables to $y = \sqrt{x}$) with a Gaussian of mean $\mu = \sqrt{\lambda}$ and standard deviation $\sigma = 0.5$, the error $g(\lambda, x)$ becomes

$$g(\lambda, x) = \frac{(\lambda - x)^3}{12x^2} - \frac{3(\lambda - x)^4}{32x^3} + ...$$

Taking the square-root of a Poisson variable thus allows approximation with a Gaussian of constant variance $\sigma^2 = 0.25$ (i.e. the transformed space is homoscedastic) and reduces the approximation error by a factor of 2 at large x.

References

[1] A. Bhattacharyya. On a Measure of Divergence Between Two Statistical Populations Defined by their Probability Distributions. *Bull. Calcutta Math. Soc.*, 35:99–109, 1943.

[2] A. C. Evans, N. A. Thacker and J. E. W. Mayhew. A Practical View-Based 3D Object Recognition System. In *Proc. Third International Conference on Artificial Neural Networks*, pages 6–15, 1993.

[3] A. Rosenfeld. Recognizing Unexpected Objects: a Proposed Approach. *International Journal of Pattern Recognition and Artificial Intelligence*, 1(1):71–84, 1987.

[4] H. Bulthoff and S. Edelman. Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proc. Nat. Aca of Sci.*, 89:60–64, 1992.

[5] D. Lowe. Object Recognition from Local Scale-Invariant Features. *Proc. IEEE International Conference on Computer Vision, Greece*, 2:1150–1157, 1999.

[6] H. Ling and K. Okada. Diffusion Distance for Histogram Comparison. *Proc. IEEE Computer Society Conference on Computer Visionand Pattern Recognition*, 1:246–253, 2006.

[7] H. Ling and K. Okada. An Efficient Earth Mover's Distance Algorithm for Robust Histogram Comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):840–853, 2007.

[8] J. J. Koenderink and A. J. Van Doorn. Representation of Local Geometry in the Visual System. *Biological Cybernetics*, 55:367–375, 1987.

[9] K. Mikolajczyk, A. Zisserman and C. Schmid. Shape Recognition with Edge-Based Features. *Proc. British Machine Vision Conference*, pages 779–788, 2003.

[10] K. Mikolajczyk and C. Schmid. A Performance Evaluation of Local Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[11] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Company, 1982.

[12] N. A., Thacker, P. A. Riocreux and R. B. Yates. Assessing the Completeness Properties of Pairwise Geometric Histograms. *Image and Vision Computing*, 13(5):423–429, June 1995.

[13] S. Belongie, J. Malik, and J. Puzicha. Shape Matching and Object Recognition Using Shape Contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2006.

[14] S.Coupe and N.A.Thacker. Quantitative verification of projected views using a power law model of feature detection. *Proc. CRV, Canada.*, 2008.

[15] M. Tarr and S. Pinker. Mental rotation and orientation dependence in shape recognition. *Cog. Psy.*, 28(21):233–282, 1989.