An Evaluation of Shape Descriptors for Image Retrieval in Human Pose Estimation

Phil Tresadern
CRHPR
University of Salford
Salford M6 6PU, UK
p.tresadern@salford.ac.uk

Ian Reid
Active Vision Lab
University of Oxford
Oxford OX1 3PJ, UK
ian@robots.ox.ac.uk

Abstract

This paper presents an empirical comparison of several shape representations in order to search a database of training examples (silhouettes) for the task of human pose estimation. In particular, we compare the Discrete Cosine Transform (DCT), Lipschitz embeddings and the Histogram of Shape Contexts that has previously demonstrated some success in this task. Our results suggest that a simple linear transformation of the image (such as the DCT) is as effective as the more complex, non-linear methods.

1 Introduction

Due to the rapid increase in affordable secondary storage over the last few years, it is becoming increasingly important to develop systems that retrieve data based on *content* rather than annotating the data by hand. This has led to the growth of interest in shape matching and retrieval algorithms with applications including searching the Web (*e.g.* Google Images) and more specific fields such as trademark enforcement. Since it is typically infeasible to use the raw, high-dimensional image to describe the data, D features are computed that retain the most informative data in the image. This dimensionality reduction provides three major benefits:

- **Lower storage requirements:** each image is reduced to a compact feature vector.
- **Increased efficiency:** the training data can be processed more rapidly.
- **Reduced sensitivity to noise:** features capture the most informative shape characteristics whilst ignoring irrelevant details.

In this work, we compare three shape representations that reduce the dimensionality of training images for the purpose of image retrieval in human pose estimation. In particular, we compare the recently proposed Histogram of Shape Contexts [1] with two simpler descriptors, namely the Discrete Cosine Transform (DCT) and Lipschitz embeddings. Although the success of the Histogram of Shape Contexts for recovering human pose was demonstrated within a sparse regression framework [1], resulting in its adoption in other studies (*e.g.* [10]), to date no empirical evidence has been presented to support claims that this is due to the efficacy of the descriptor rather than the regressor. This work presents the first quantitative comparison to investigate this claim by comparing representations *under controlled conditions where meaningful comparisons can be made*.

BMVC 2007 doi:10.5244/C.21.65

1.1 Related Work

The range of shape descriptors available for applications such as human pose estimation from binary silhouettes is very large. However, we can argue that many representations are inappropriate for this task. Descriptors based on the topology of the occluding contour [7] change dramatically with small changes in underlying pose (*e.g.* as the subject places their hands on their hips such that 'holes' are created that modify the topology). Representations based on curvature [15] typically require a continuous (or sufficiently high resolution) contour that is rarely available in this application. Similar arguments rule out Fourier decompositions [16] and shock graphs/median axis representations [9].

Of the remaining candidates, *global* representations use every pixel to compute every feature such that a localized corruption of the input image (*e.g.* due to occlusion or shadow) induces an error in every feature. Such representations include embeddings [5], moments [8, 12, 14] and Principal Component Analysis (PCA). In contrast, *local* representations use only a subset of the image to compute each feature such that only certain features are affected by a localized error in the input image. Such representations include the recently proposed Histogram of Shape Contexts (HoSC) that has successfully been employed in human pose estimation [1]. It is this property of locality that is claimed to make such representations superior.

1.2 Paper structure

We begin in Section 2 by describing the selected shape descriptors, including a discussion of how appropriate parameters were selected for each. Section 3 describes the experimental data and how the descriptors were compared. Results are presented in Section 4.

2 Shape representation

2.1 Discrete Cosine Transform (DCT)

We begin with a form of the Discrete Cosine Transform of the $P \times Q$ image, $I(x,y)$, whereby each feature (DCT coefficient), M_{mn}, is defined by:

$$M_{mn} = \sum_x \sum_y f_m(x) I(x,y) f_n(y) \tag{1}$$

and we define

$$f_m(x) = \sqrt{\frac{1+\min(m,1)}{P}} \cos\left\{\frac{m\pi}{P} \cdot \left(x+\frac{1}{2}\right)\right\} \tag{2}$$

where $m = 0 \ldots P-1$ and $x = 0 \ldots P-1$. This transform can be an interpreted as a rotation of the vectorized image such that the Euclidean distance between feature vectors in PQ-dimensional space is equal to the sum of squared error between the original images. Using only a subset of D coefficients therefore approximates the SSE between images. Furthermore, this form of the DCT belongs to the family of *orthogonal moments* since:

$$\int f_i(x) f_j(x) dx = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{3}$$

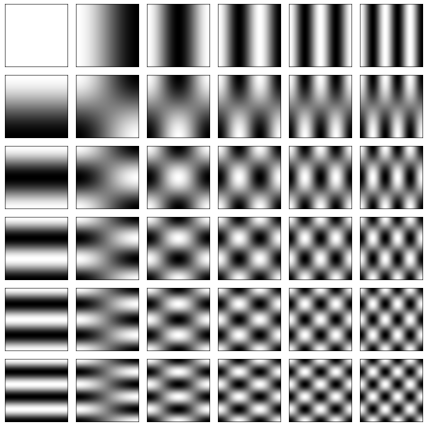


Figure 1: Filter bank equivalents (up to order 5) of DCT moment generating functions, $f_{mn}(x,y) = f_m(x)\,f_n(y)$.

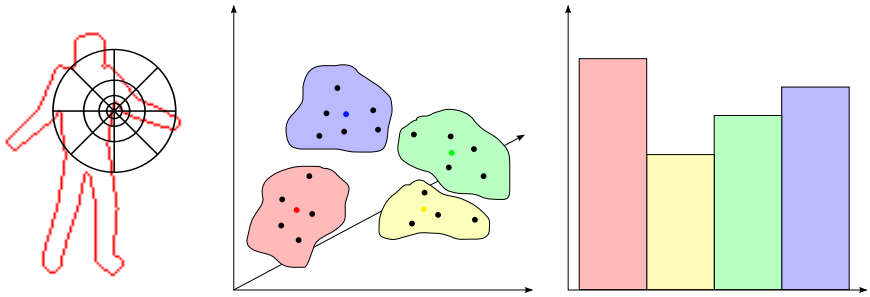
such that correlation is low between coefficients and fewer are required (compared to non-orthogonal moments) to describe the image within a given error bound.

Other transformations were also considered such as Tchebichef [8], Krawtchouk [14], geometric and Hu [6] moments in addition to PCA. Although PCA provides an optimal (in terms of capturing maximum variance) basis set over the set of images, the basis set is data-dependent and impractical to compute for the image sizes involved. Tchebichef moments were found to be qualitatively similar to the DCT, effectively providing a frequency decomposition of the image, although with slightly worse performance in the evaluation task. Krawtchouk moments (another orthogonal moment) also performed slightly worse than the DCT, possibly as a result of limited spatial support of lower order moments.

Geometric moments are seldom employed due to the concentration of 'mass' at the edges of the image (where the least informative data resides) and the lack of an intuitive distance metric between feature vectors (in contrast to orthogonal moments). Similarly, although Hu moments are popular due to their rotational invariance they are based on geometric moments and hence suffer the same shortcomings. Furthermore, only seven Hu moments are typically defined which do not capture sufficient variation in many datasets.

In order to make the comparison fair, we first undertook a number of experiments to assess the impact of various parameters [13]. These experiments suggested that:

- Although performance improved as more DCT coefficients were retained (since the distance between feature vectors more closely approximates the true SSE between images), most useful information was captured by $D \geq 64$ features.
- When ranking the database in order of similarity to the query in feature space, Euclidean distance (the most intuitive metric since it is directly related to the SSE) gave very similar performance to the Mahalanobis and Manhattan (L_1) distances.
- Feature selection heuristics such as maximum order ($\max\{m,n\}$), order ($m+n$) and RMS value all gave similar results whilst variance was a poor indicator of feature information. More complex feature selection is beyond the scope of this work.



(a) (b) (c)

Figure 2: Overview of HoSC descriptor: (a) Each contour point is assigned a high-dimensional 'Shape Context' based on the local distribution of other contour points; (b) Shape Contexts from all database examples are clustered to generate D cluster centres (codebook vectors); (c) A normalized histogram is generated for each example based on the distribution of cluster centres voted into by the Shape Contexts of its contour points.

2.2 Lipschitz embeddings

The second global representation we consider is the Lipschitz embedding [5], whereby an image is represented by the vector of distances from the query image to D 'pivot' exemplars and has recently demonstrated success in hand tracking applications [3]. More specifically, we embed each image by extracting its contour points and computing its (asymmetric) chamfer distance from the ith pivot examplar to give the ith element of the feature vector. Intuitively, images that are close together in image space have similar distances to the pivot examples and therefore have similar feature vectors. However, selecting pivots from the same region of space results in highly correlated (*i.e.* redundant) features that may degrade performance.

Experiments to investigate the effect of various parameters [13] suggested that:

- Most information for this dataset was captured using $D \geq 100$ features (pivot exemplars).
- Due to the non-linear nature of the Lipschitz embedding, it is difficult to identify an intuitive distance metric between two feature vectors. However, using the Mahalanobis distance resulted in a noticeable improvement over the Euclidean and Manhattan metrics.
- No significant difference in performance was observed over 100 randomly selected sets of exemplars although a more intelligent approach to feature selection was recently investigated using Boosting [2].

2.3 Histogram of Shape Contexts (HoSC)

Our final selected shape descriptor is the Histogram of Shape Contexts, suggested by Agarwal and Triggs [1], and demonstrated using silhouettes of the human body. In this representation (see Figure 2), every point along the contour of the silhouette is assigned a histogram (known as its Shape Context [4]) representing the distribution of other contour

Figure 3: In this example, both the angel and the demon are composed of identical contour segments such that their histograms become indistinguishable as the spatial extent (*i.e*. the radius) of the shape context vector approaches zero. Note that *exact* tesselation is not required for very different silhouettes to result in very similar feature vectors.

points in a local neighbourhood (defined by the Shape Context 'radius'). Having computed the Shape Context for all contour points on all silhouettes in the database, D Shape Contexts are then selected at random and used as initial centres in a k-means clustering scheme. Following clustering, the updated cluster centres are used as a vector quantization 'codebook' in order to assign each contour point on a given silhouette to a cluster. A histogram over cluster assignments then forms the feature vector for a given silhouette. This histogram should be normalized with respect to the number of contour points to make the descriptor scale-invariant. Furthermore, in order to reduce quantization effects, 'soft' voting allows each contour point to vote into more than one bin.

It is suggested that this descriptor may be superior due to its locality – corrupting a small region of the silhouette should modify only a few features, in contrast to the DCT and Lipschitz embeddings where the whole silhouette contributes to every feature. However, we note that: (i) in most cases the corruption of the silhouette (*e.g*. due to shadows or occlusion) results in an increase or decrease in the number of contour points such that normalizing the histogram then affects *every* bin; (ii) typical distance metrics (*e.g*. Euclidean distance, Bhattacharyya coefficient) do not exploit this locality in any beneficial way; (iii) no explicit distinction is made between the interior and exterior of the silhouette, thus discarding potentially valuable information (see Figure 3).

These concerns provided the motivation behind comparing the Histogram of Shape Contexts to other descriptors in order to quantify any benefit gained from the substantial increase in computational complexity. As with the other descriptors, a basic analysis of the parameters [13] suggested that:

- Again, most information was captured by $D \geq 64$ features (codebook vectors).
- The use of intuitive distance metrics for histograms (*e.g*. Bhattacharyya distance) did not significantly improve performance over other (less correct) metrics such as the Manhattan and Euclidean distance (this has previously been attributed to 'soft' voting [1]).
- Since codebook vectors are typically well distributed after clustering, performance was largely insensitive to their initial random selection as evaluated over 100 trials.

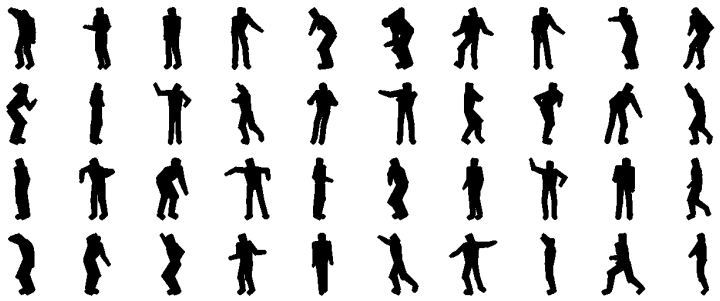


Figure 4: Example silhouettes from the synthetic dataset.

- Performance was stable for any sensible Shape Context 'radius' of at least the mean distance between all pairs of contour points.
- Although we used 12 angular bins (a common value), performance is stable for any value above 8. Performance was largely invariant to the number of radial bins.
- The use of 'soft' voting (as advised in [1]) to avoid quantization effects provided a small benefit when each contour point voted into > 4 bins.

3 Method

In order to evaluate the selected shape descriptors, we used motion capture data (available at the time of printing from `http://mocap.cs.cmu.edu`) to generate N=10000 128×128 binary silhouettes of a human body model (Figure 4). This training set included synthetic silhouettes from several different 'exercise' motions generated from 4 camera locations equally spaced from 0° to 90° in azimuth.

In addition to the training data, an additional 250 silhouettes were generated from synthetic data to test the retrieval performance of the shape descriptors. Furthermore, 40 real test images were obtained by background subtraction of several sequences of a subject undertaking exercise motions similar to those in the training data.

For the purposes of this evaluation, all images were normalized by translating and scaling the silhouette such that it lay within the central 90% of the image. We also assumed that the subject was upright in the image to avoid any need for rotation invariance; any exceptions to this rule (*e.g.* handstands, cartwheels) were explicitly modelled in the dataset. All silhouettes were then reduced to a feature vector of $D = 100$ dimensions using each of the proposed descriptors.

Silhouettes generated from synthetic data were automatically labelled with the image projections of the joint centres since these values were directly available. For silhouettes obtained from real sequences, the image projections of joint centres were labelled manually using the mouse in order to evaluate performance.

Like many other studies, we employ silhouettes since they are readily obtained from image data by background subtraction and are relatively invariant to clothing and lighting. However, they are generally restricted to scenes with a static camera and known background, and useful image data (*e.g.* internal edges) are discarded.

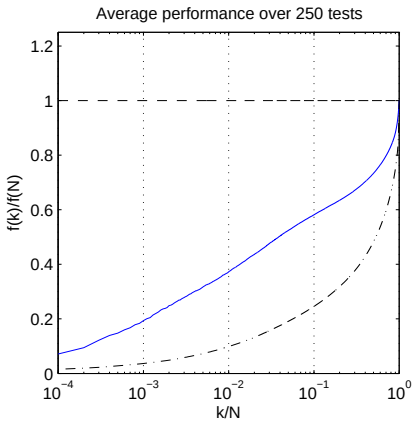


Figure 5: Example graph of k/N against $f(k)/f(N)$. For comparison, the dashed line at unity indicates the average curve produced by random ordering whilst the dash-dot curve indicates the best possible ranking where distance in image space correlates perfectly with distance in pose space.

3.1 Evaluation method

Image retrieval tasks typically require *classification* of the query input such that stored examples of the same class are returned. Recovered exemplars are therefore classed as positive or negative and evaluation tools such as the Receiver Operating Characteristic (ROC) curve and Precision-Recall curve may be used to compare retrieval accuracy between different shape descriptors.

In the context of human pose estimation, however, exemplars cannot be classified into 'positives' and 'negatives' since the underlying pose space is continuous. Therefore, we use the sum of squared errors between corresponding joint centre projections[1] in the image to compute the distance, $d(x_i, x_q)$, in pose space between each training example, x_i, and a query, x_q. Given a query silhouette, we rank the training data in order of similarity to the query as quantified by the chosen shape descriptor, denoting the index of the closest training example by $r(1)$ and the furthest by $r(N)$. We then generate a curve, $f(k)$:

$$f(k) = \frac{\sum_{j=1}^{k} d(x_{r(j)}, x_q)}{k}, \tag{4}$$

indicating the mean distance to the query of the k highest ranking training examples for $k = 1 \ldots N$. For a qualitative performance evaluation, we compare the normalized curve of k/N against $f(k)/f(N)$ in addition to the corresponding curves for the expected performance of a random ranking of the training data (*i.e.* unity) and for the best possible ranking, as shown in Figure 5. Each curve can be interpreted as a measure of correlation between distance in state space and distance in feature space – high correlation (desirable) produces a 'low' curve whereas low correlation produces a 'high' curve.

[1] Using *projected* joint centres rather than their full 3D position avoids many (though not all) problems associated with 'kinematic flip' ambiguities [11] where very different poses give rise to very similar projected joint centres.

Figure 6: Four test datasets: (a) clean silhouettes; (b) with added noise; (c) with lower quarter removed; (d) real silhouettes manifesting some segmentation error.

4 Results

We compared the three selected shape descriptors using four test datasets (Figure 6) containing silhouettes that were: (i) perfect; (ii) noisy; (iii) partially occluded; (iv) real.

We begin by comparing the three methods for clean data (Figure 6a) taken directly from the synthetic dataset. Figure 7a shows that, although Lipschitz embeddings perform slightly worse than the other descriptors, accuracy is similar for all three representations.

To create a noisy data-set, we corrupted the clean test silhouettes with Gaussian noise along the contour (Figure 6b). Such corruption typically results from segmentation errors at the boundaries and compression artefacts. From Figure 7b, we see that performance is largely unchanged by the added noise, with the exception that DCT coefficients marginally outperform the Histogram of Shape Contexts. This may be explained by the fact that lower order DCT coefficients (as used in this case) encode the lower frequencies within the image and therefore suppress noise. Again, Lipschitz embeddings do not perform as well as the other two methods.

In order to simulate occluded data, we removed the bottom quarter of each test silhouette and renormalized, as if the subject had been obscured from approximately knee-level down (Figure 6c). Although this is a relatively crude approach, it presents each method with data that is somewhat different from the training data yet is typical in real life applications. Figure 7c shows that the Histogram of Shape Contexts performs well for small k (approximately the top 1% of the data) but is out-performed for higher k by the DCT. Lipschitz embeddings are again typically out-performed by the other two methods.

For the final experiment, we use real silhouettes from a ‘starjumps’ sequence (Figure 6d), obtained via background subtraction and with projected joint centres labelled by hand. Due to the limited number of test images, the curves in Figure 7d are slightly noisier but suggest that DCT coefficients significantly outperform both Histogram of Shape Contexts and Lipschitz embeddings. More specifically, the Histogram of Shape Contexts and Lipschitz embeddings have perform similar to a random ranking for this data-set. This is a surprising and interesting result, particularly since this is arguably the most important test set of the four. It may be questioned whether the normalization procedure employed in this experiment might favour one method over another. However, the test silhouettes show little corruption that would have a significant effect on this process.

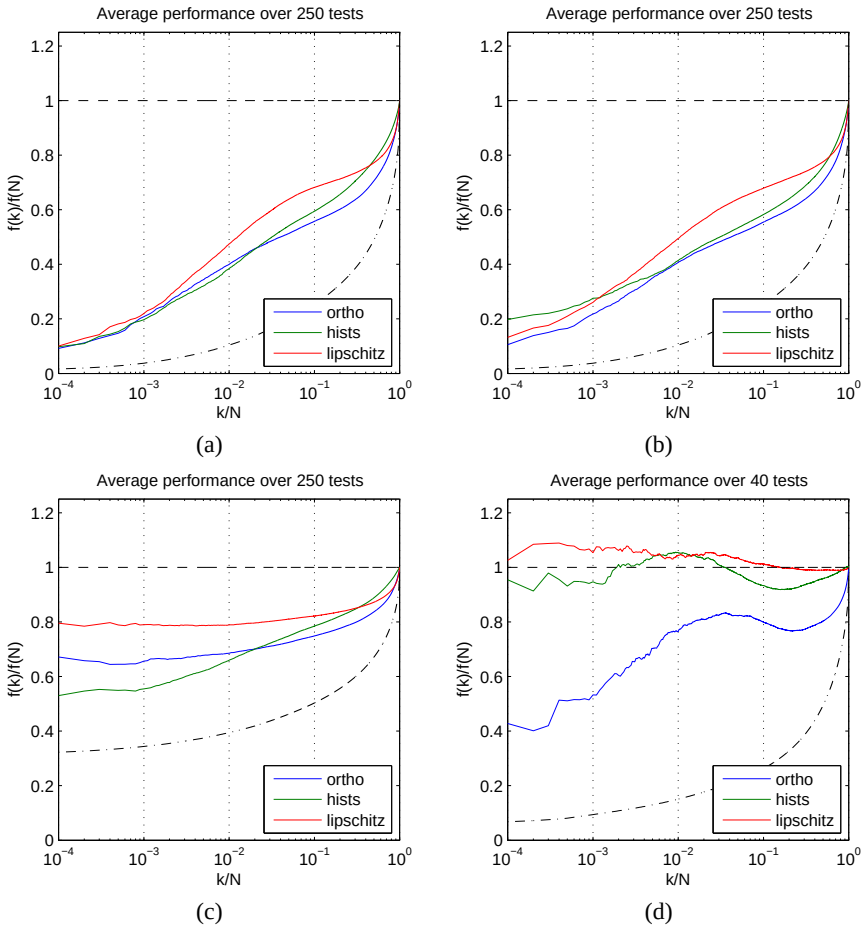


Figure 7: Results for (a) clean data; (b) noisy data; (c) occluded data; (d) real data. Curves correspond to DCT coefficients (*ortho*), Histogram of Shape Contexts (*hists*) and Lipschitz embeddings (*lipschitz*)

5 Conclusion

We have presented a comparison of three shape descriptors for the application of human pose estimation from binary silhouettes. In particular, we compare two straightforward and established methods (the DCT and Lipschitz embeddings) against the recently proposed Histogram of Shape Contexts (HoSC), a 'local' descriptor that is claimed to be superior to 'global' methods. However, despite its computational complexity, our results suggest that the HoSC offers little (if any) benefit over the alternative, simpler methods.

Although it has not escaped our attention that some of our results appear to contradict those that have appeared in previous works, we note that these studies often employed a limited number of training images [1] or more a complex matching process [2]. To the best of our knowledge, this study is the first to evaluate such descriptors under controlled conditions where meaningful comparisons can be made.

References

[1] A. Agarwal and B. Triggs. Recovering 3D human pose from monocular images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(1):1–15, January 2006.

[2] V. Athitsos, J. Alon, S. Sclaroff, and G. Kollios. BoostMap : A method for efficient approximate similarity rankings. In *Proc. 22nd IEEE Conf. on Comp. Vis. and Patt. Rec.*, volume 2, pages 268–275, 2004.

[3] V. Athitsos and S. Sclaroff. Estimating 3D hand pose from a cluttered image. In *Proc. 21st IEEE Conf. on Comp. Vis. and Patt. Rec.*, volume 2, pages 432–442, 2003.

[4] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):509–522, April 2002.

[5] G. R. Hjaltason and H. Samet. Properties of embedding methods for similarity searching in metric spaces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(5):530–549, May 2003.

[6] M. K. Hu. Visual pattern recognition by moment invariants. *IRE Trans. Inform. Theory*, 8:179–187, February 1962.

[7] L. J. Latecki and R. Lakamper. Convexity rule for shape decomposition based on discrete contour evolution. *Comput. Vis. Image Und.*, 73(3):441–454, March 1999.

[8] R. Mukundan, S. H. Ong, and P. A. Lee. Image analysis by Tchebichef moments. *IEEE Trans. Image Process.*, 10(9):1357–1364, September 2001.

[9] K. Siddiqi, A. Shokoufandeh, S. J. Dickinson, and S. W. Zucker. Shock graphs and shape matching. *Int. J. Comput. Vis.*, 35(1):13–32, November 1999.

[10] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative density propagation for 3D human motion estimation. In *Proc. 23nd IEEE Conf. on Comp. Vis. and Patt. Rec.*, volume 1, pages 390–397, 2005.

[11] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3D human tracking. In *Proc. 21st IEEE Conf. on Comp. Vis. and Patt. Rec.*, volume 1, pages 69–76, 2003.

[12] M. R. Teague. Image analysis via the general theory of moments. *J. Opt. Soc. Am.*, 70:920–930, August 1980.

[13] P. Tresadern. *Visual Analysis of Articulated Motion*. PhD thesis, University of Oxford, October 2006.

[14] P.-T. Yap, R. Paramesran, and S.-H. Ong. Image analysis by Krawtchouk moments. *IEEE Trans. Image Process.*, 12(11):1367–1377, November 2003.

[15] D. S. Zhang and G. Lu. A comparative study of curvature scale space and Fourier descriptors. *J. Vis. Commun. Image R.*, 14(1):41–60, March 2003.

[16] D. S. Zhang and G. Lu. Study and evaluation of different Fourier methods for image retrieval. *Image Vision Comput.*, 23(1):33–49, January 2005.