Simple canonical views

Peter Hall and Martin Owen
Department of Computer Science
University of Bath
Bath, BA2 7AY
pmh | cspmjo @cs.bath.ac.uk

Abstract

This paper demonstrates how to determine canonical views of objects in a way that is simple, robust and versatile. Common parlance loosely defines canonical views as the “front”, “side”, and “top” views of an object. Our approach determines these views for objects whether represented by images or three-dimensional points; neither image segmentation nor model analysis is required. It is easy to introduce constraints so that other views can be determined, as desired. We explain our method and compare it qualitatively to alternatives.

1 Introduction

Canonical views of objects are of interest to both the Computer Vision [2, 9, 11] and the biological vision [1] communities, indeed the term “canonical view” seems to have been coined in 1981 by the latter community [7]. Although literature is unable to agree on a common definition of what a canonical view is, they share two points in common: (1) all are premised on the idea that “view stability” of some kind is important and a complex algorithm is needed to determine these views, and (2) none determine views corresponding to the lay definition that canonical views are the “front”, “side”, and “top” views.

We differ on both points – the contribution of this paper is to provide a method that automatically determines the front, side, and top views in a way that is simple, robust, and versatile. Our motivations were both philosophical and pragmatic. From a purely philosophical stance it is important to conform to commonly held definitions rather than to stipulate definitions of one’s own; breach of this principle has given rise to disagreement as to what a canonical view actually is, typically justified by appeal to criteria that are related to efficiency of representation. Pragmatically, the lay definition of canonical views is of value, and is often an efficient representation. It is widely used in technical drawings produced by architects and engineers, in medical photographs and drawings, in photographs of museum pieces, and in paintings produced by children and in many schools of Art spread throughout all history and every continent.

Empirical evidence in support of the intuition that humans seem to prefer the front, side, and top set of views can be found in the work of Min *et al.* [12] where people were asked to sketch views of 3D objects and found,

> the most frequently chosen views were not the characteristic views predicted by perceptual psychology, but instead ones that were simpler to draw (i.e. front, side and top views).

Also the findings of Perrett and Harries [8] suggest that people often prefer views aligned to the principal axes of three-dimensional objects. Unfortunately principle axes do not always provide canonical viewing directions as we understand them, most obviously for objects that exhibit little of no symmetry — an L-shaped building, for example. Despite this, the notion that axes of symmetry provide a canonical basis set is a useful idea; Hong *et al.* [6] make use of such basis sets in a variety of ways, including reconstruction of objects from a single image.

Yet much of the Computer Vision literature eschews the common definition of canonical views (equivalently; canonical viewing directions or canonical reference frames). Freeman [3] and also Weinshall and Werman [11] argue in favour of the most likely views one the grounds that these are the most stable, in the sense that small changes from them make the least difference. Similarly, and more recently, Peters *et al.* [9], prefer views that cover the widest area of the viewing sphere. Denton *et al.* [2] state that canonical views are those that most efficiently characterise a set of views, and use distance measures between views to determine the canonical set. It is clear that the definition of what constitutes a "canonical view" depends upon the applications that the researcher has in mind; some kind of stable view is preferred because it is assumed reconstruction applications benefit from such views.

Our definition of canonical views is based on observation; we define *canonical views to be the most unique amongst all views, subject to orthogonality constraints*. In other words, canonical views are the least likely views and tend to be highly unstable.

Our method for determining canonical views, finding the rarest datum amongst a set, is simple statistics. Because we are looking for outliers the system tends to be robust to small variations. The system is versatile not only because we can generate datum from images or projections of point sets, but because we can associate viewing directions and so impose geometric constraints. Furthermore we can change our mind about what a canonical view is and choose the most likely datum, that emulates the broad features of that section of Computer Vision literature which we ostensibly disagree with.

2 Finding Canonical Views

Our problem is to determine these canonical views, given set of images from the view sphere, or part of the view sphere. We assume invariant internal camera parameters, and that the distance between the object and camera changes little. We define a set of canonical views as the least likely set of views, subject to constraints. The constraints we choose are that the view directions should be mutually orthogonal. We hope to obtain the front, side, and top views.

We begin by normalising images to allow for lighting variations; we ensure pixel values sum to unity; we perform no other preprocessing. As is common, each image is treated as a high-dimensional vector in which each pixel value is a vector element. Colour pictures are converted to greyscale beforehand.

Given a set of n data (images) $\mathbf{x}_i$, each with an associated viewing direction $\mathbf{f}_i$ we build an eigenmodel $\Omega = (n, \mu, \mathbf{U}, \Lambda)$ to describe their distribution. The high volume of data (up to 2500 images) and high dimension of each vector forces us to use an incremental approach when constructing the eigenmodel [4]. Typically we retain about 97% of the eigenenergy, leaving a low dimensional representation of each images, typical 10 or 20 eigenvectors remain.

We then measure the Mahalanobis distance

$$m_i = (\mathbf{x}_i - \mu)^T \mathbf{U}^T \Lambda^{-1} \mathbf{U}(\mathbf{x}_i - \mu)$$

of each datum. The image with the largest Mahalanobis distance is the least likely, so it is selected as the initial canonical view $\mathbf{y}_1$, and a corresponding view direction $\mathbf{g}_1$

$$\begin{aligned} k &= \underset{i}{\text{argmax}}\ m_i & (1) \\ \mathbf{y}_1 &= \mathbf{x}_k & (2) \\ \mathbf{g}_1 &= \mathbf{f}_k & (3) \end{aligned}$$

Having obtained an initial canonical view we filter out all those that were not captured from a (nearly) orthogonal vantage point and select the least likely image and view direction from amongst those that remain:

$$\begin{aligned} \mathscr{J} &= \{i : |\mathbf{g}_1^T \mathbf{f}_i| < \varepsilon\} & (4) \\ k &= \underset{i}{\text{argmax}}\ m_i \text{ subject to } i \in J & (5) \\ \mathbf{y}_2 &= \mathbf{x}_k & (6) \\ \mathbf{g}_2 &= \mathbf{f}_k & (7) \end{aligned}$$

Where ε is typically a measure of machine accuracy, but could be increased to relax the orthogonality constraint. We continue this procedure to obtain the final canonical view:

$$\begin{aligned} \mathscr{J} &\leftarrow \mathscr{J} \cap \{i : |\mathbf{g}_2^T \mathbf{f}_i| < \varepsilon\} & (8) \\ k &= \underset{i}{\text{argmax}}\ m_i \text{ subject to } i \in J & (9) \\ \mathbf{y}_2 &= \mathbf{x}_k & (10) \\ \mathbf{g}_2 &= \mathbf{f}_k & (11) \end{aligned}$$

Although this simple approach has proven very effective it can sometimes produce the *opposite* view to that desired, the back instead of the front, for example. This behaviour is easily remedied by selecting the opposite point of view too, provided it lies within the data set.

Figure 1 shows a typical surface constructed by scaling the unit viewing hemisphere: $\mathbf{p}_i = M_i \mathbf{f}_i$. It is clear that the points lie on a convoluted surface, which we will call a "Mahalanobis surface". Canonical views lie on convex lobes of the Mahalanobis surface. Figure 2 shows the corresponding canonical photographs, with views from opposite directions appended. Figure 3 is a small gallery of results obtained from various objects.

3 Variations on a canonical theme

Here we demonstrate the versatility of the method by considering a few of the many variations that could exist.

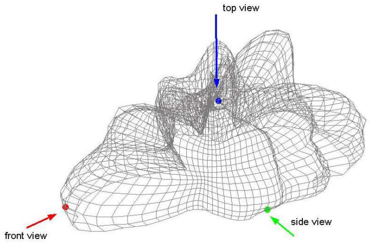

Figure 1: A typical Mahalanobis surface, in this case for the set of photographs represented in Figure 2. Canonical view directions views are shown in colours, red, green, blue.

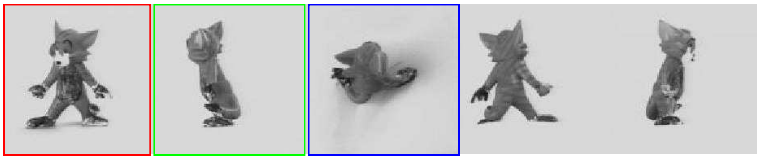

Figure 2: Canonical views identified from 2500 images, using the Mahalanobis surface in Figure 1; the three left-most pictures correspond the identified *front*, *side*, and *top* view directions in that order. The two right-most pictures have been appended as opposite views. Images courtesy of Peters *et al.* [9].

3.1 Removing the explicit orthogonality constraint

Some readers might object to the forcing of orthogonality between canonical views. This issue is addressed by processing the Mahalanobis surface. We have already observed that this surface is highly convoluted, and that the canonical views thus far lie on convex sections. Ideally we would like to partition this surface into convex and concave components so as to create disconnected components. The number of canonical views would then be equal to the number of connected components, and the view with the largest Mahalanobis distance in each component would be taken to be a canonical view. Furthermore we might hope to broaden the definition of “canonical view” to take views from the concave partitions; the views so obtained could then be compared to the views obtained by other methods because these would be the most likely views.

Unfortunately the Mahalanobis surface is small, noisy, and difficult to segment into convex and concave parts. Furthermore, the concave parts form a single connected segment; the convex components appear as “islands” within a convex sea. We therefore adopt a slightly different approach, as follows.

We consider the Mahalanobis surface as a height field of two independent variables, (θ,ϕ). This is robustly filtered to identify local minima, maxima, and saddle points. We then look for groups of such critical points that are as widely spread as possible; if maxima are used we can closely reproduce the front, side, and top views results obtained in the standard version of our approach; if minima are used as critical points we obtain a set of the most likely views.

The robust filter processes each direction independently, so we need only consider

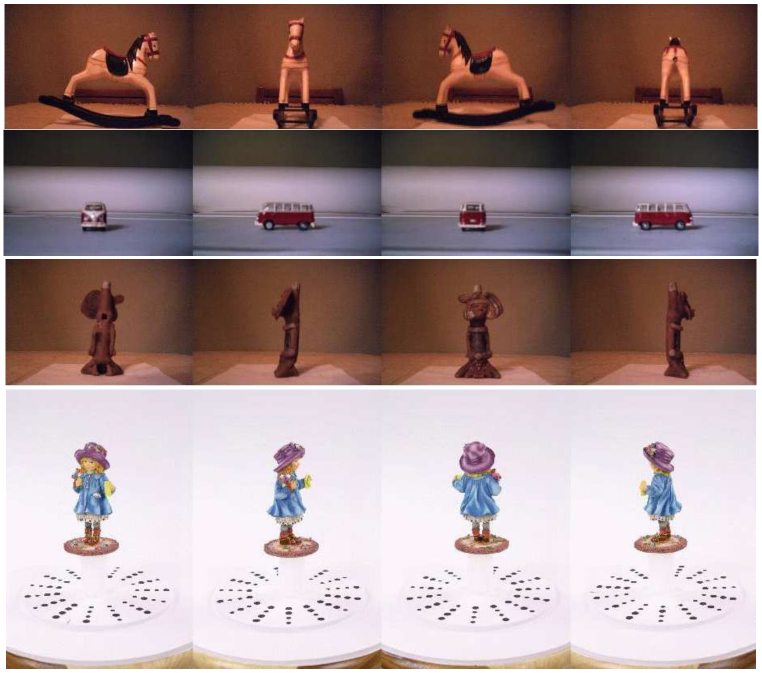

Figure 3: A gallery of results for various objects. The toy rocking-horse, the toy campervan and the small clay statue were all obtained on a low-quality camera with the object being turned by hand, using a table-top grid as a guide; twelve images per object, with a single rotational degree of freedom. The doll images are courtesy of Adam Baumberg of 3D-SOM, and were acquired with a high-quality camera, the object being on a turn-table; 15 images were used to decided canonical views. Some images taken from above and below the doll were provided but not used.

a one dimensional problem. Given a line of constant θ, say, we look for all minima and maxima using a morphological filter based loosely on sieves [5]. An array element at location i is a maximum of integer half-width w if $h(\theta,i) \geq h(\theta,j)$ for all $j \in [i-w, i+w]$. We define a minimum similarly. Thus we construct a scale-dependent map of extrema, $m(\theta,\phi,w) = 1$ at maxima, -1 at minima and 0 elsewhere in row-by-row fashion.

An appropriate scale is determined using the principle that salient features are stable over scale; this was based on the maximally stable regions advocated by Obdržálek and Matas [10]. Given a set of maps $m(\theta,\phi,w)$ the squared distance between adjacent maps is $d(w,w+1) = |m(\theta,\phi,w+1) - m(\theta,\phi,w)|^2$. The number of extrema at scale w is $n(w)$. We use an exhaustive search to minimise $d(w,w+1)/n(w)$, and hence an appropriate w, and hence a set of stable extrema. Having obtained stable maps in each direction, $m_\theta(\theta,\phi)$ and $m_\phi(\theta,\phi)$ it is easy to find minima, maxima, and saddle points, $\max(\theta,\phi) = (m_\theta(\theta,\phi) > 0) \wedge (m_\phi(\theta,\phi) > 0)$, for instance.

To determine a set of N canonical views we consider all $N - tuples$ of the critical points (minima, maxima or saddle points) to find the tuple whose viewing directions are the most spread. The definition of "spread" we used is designed to coincide with the orthogonality constraint previously used without imposing it directly. Suppose $\mathbf{X} =$

$[\mathbf{x}_1 \dots \mathbf{x} - N]$ is an N-tuple of viewing directions, with each $\mathbf{x}_i \in \Re^3$. We define the *spread* of these direction vectors as proportional to the mean of all inner products:

$$s(\mathbf{X}) \quad = \quad 1 - \frac{1}{9}[111](|\mathbf{X}'\mathbf{X} - \mathbf{I}|)\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \tag{12}$$

in which $|\mathbf{Y}|$ is a matrix with all elements their absolute value and $\mathbf{I}$ is the (3×3) identify matrix. We search for the $N - tuple$ to maximise $s(.)$. Just as when an orthogonality constraint was explicitly enforced the system can produce "back" views rather than "front" views; again just as easily handled by arguing that a viewing direction can be looked along in two ways, and so picking opposite views that exist.

Figure 4 shows that this method is capable of producing canonical views that closely conform to our definition. Unfortunately the generality brought by this modified definition can lead to a slight degradation on performance; as demonstrated by the results for a toy dwarf, shown in Figure 5.

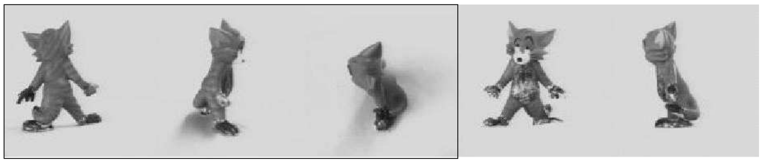

Figure 4: Canonical views of a toy cat determined by analysis of the Mahalanobis surface, subject to loose geometric constraints. The initial 3-tuple of pictures chosen are boxed, opposite views that exist are appended; compare this to Figure 2.

3.2 Working with point sets

As well as sets of images our method can be used to generate canonical viewing directions for 3D computer models. Using only point sets a sphere of viewing directions is chosen and the points projected down to a plane using an affine camera. An eigenmodel is then built from the projected points and canonical views can be chosen from the eigenmodel just as for images, see Figure 6. Here we also see the process working well on a non-symmetrical object, namely the violin.

4 Comparison with previous work

Recall that the previous literature defined canonical views as being those views that are in some sense the most-likely or most-stable. The ability to choose maxima, minima or saddles as critical points, introduced in Subsection 3.1, allows us to conform to or deviate from that general line, as we choose.

Choosing to conform to the general definition of *most-likely* views, Figure 7 compares canonical views of the toy cat using three methods. The views determined by our method used minima of the Mahalanobis surface as critical points. The Peter's *et al.*reported canonical views for the toy in [9]; views we have reproduced here. The method of Denton *et al.* [2] requires the solution of a graph-cut problem subject to (non-geometric)

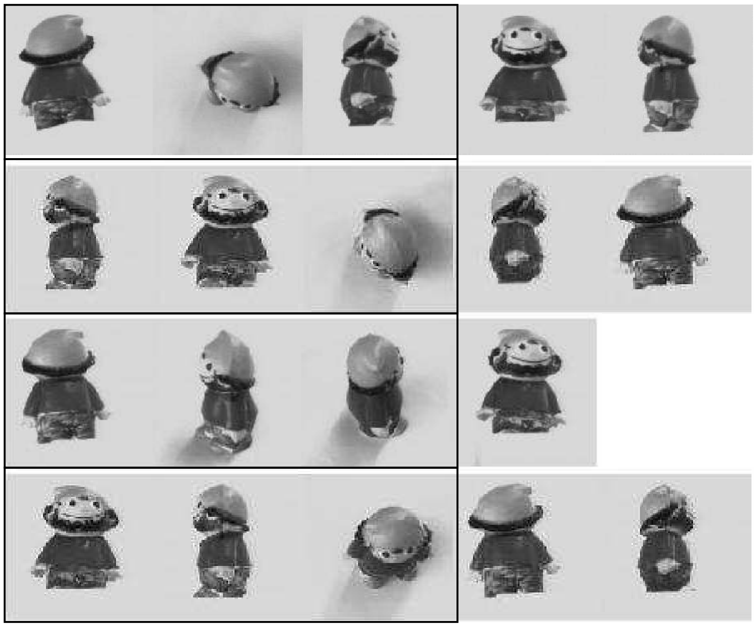

Figure 5: Canonical views of a set of 2500 images using 3-tuples, row by row; up to two images may be appended to the right each row because canonical views can occasionally locate the “back” instead of “front” view, as explained in the body of the text; initially chosen images are boxed. Top row shows canonical views based on the explicit enforcement of orthogonality; second row the maxima of the Mahalanobis surface for the object; middle-row are views based on minima; bottom-row uses saddle-point images. Images courtesy of Peters *et al.* [9].

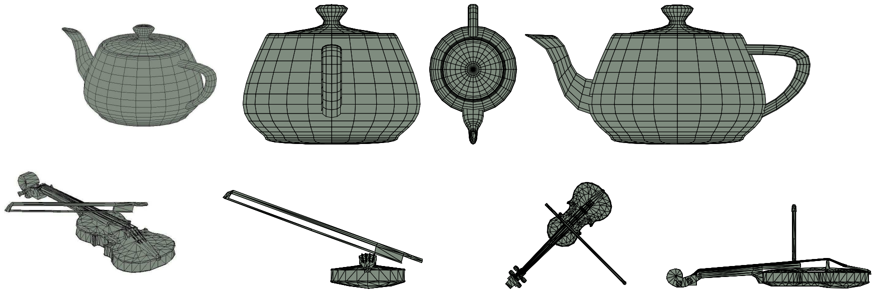

Figure 6: A teapot and violin objects (far left) with canonical views of the teapot and violin objects chosen from 2500 views of point sets.

constraints. For a large number of images (2500 in the Tom set) there are 2500^2 edges in the graph, making a graph-cut approach very expensive. This sets a practical limit on the graph-cut approach, which forces us to take a single ring of views with a fixed altitude; Denton *et al.* [2] only provide canonical views at a single altitude also.

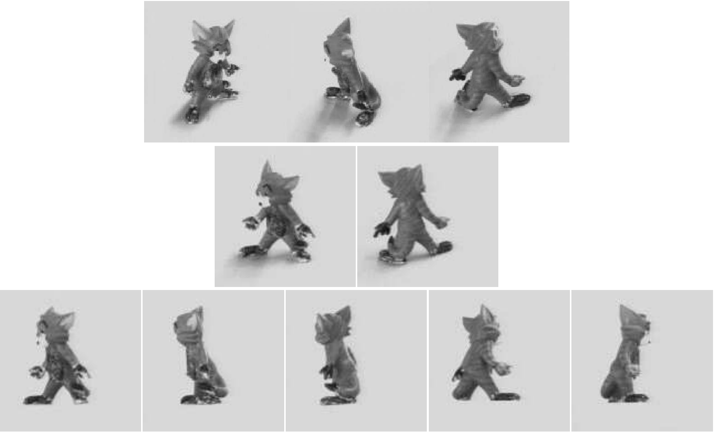

Figure 7: Canonical views of a toy cat, defined now as the most-likely view. Top row: our method of Subsection 3.1 using minima of the Mahalanobis surface. Middle row: as reported by Peters *et al.* [9]. Bottom row: as determined by our implementation of Denton *et al.* [2]

5 Discussion and Concluding Remarks

We have introduced a method to determine canonical views of objects. Its novelty arises from the fact it is able to determine the front, side, and top views of an object that conform to the lay understanding of characteristic views. We claim simplicity, robustness and versatility as advantages.

Simplicity is an uncontroversial claim, the basic algorithm is surprisingly straight forward. The slightly more complex but more general algorithm can be simplified by not filtering for stable extrema but choosing all local extrema instead. In our experience the N-tuple of views chosen as canonical remains the same, but the search takes longer. This is because the time taken to find an N tuple from a set of k elements rises in proportion to k^N; filtering reduced k. One possible way forward would be to introduce a more efficient search, perhaps by maximising spread using branch-and-bound.

Robustness is demonstrated by the top three rows of Figure 3. Recall the object in these were turned by hand over a radial-polar grid placed on the table. This led to noticeable translations of the object, both to and fro and left and right; these variations are most obvious when the images are quickly flicked through. The movement is not too great, but sufficient to show some tolerance and so raise questions such as *how tolerant?* that we have not answered here. The camera was not mounted, so that some photographs are blurred; the photograph showing the left flank of the rocking-horse is a very poor photograph (most evident when seen at full size). Again the question of degree of tolerance arises. Also, in these hand acquired images, lighting variations are considerable, especially in the camper van set where sharp specular highlights, from the van's trimmings, appear and disappear. The simple normalisation we used compensated for these; whether more sophisticated pre-processing is required is an unanswered question.

Versatility has been shown in several ways. First we have shown that the algorithm operates equally well whether images or 3D models are used as source data. Second we have shown that we can produce nearly the same results using either by explicitly enforcing orthogonality, or else by choosing maxima of the Mahalanobis surface and maximising the spread of view directions. Thirdly we have shown that by choosing minima instead and them maximising we choose views that are conform more closely to the general understanding of canonical views as advocated by the technical literature. The conclusion is that the Mahalanobis surface plays an important unifying role in determining canonical views.

Choosing views by processing the Mahalanobis surface does not always produce exactly the same set of views as explicit enforcement of orthogonality. Nonetheless, the fact the two sets of views are close approximations of one another is important, for it shows that the lay understanding of canonical views is not arbitrary but can be explained by reference to an objectively measurable artifact; the convex lobes of Mahalanobis surface tend to be in orthogonal directions. Hence we can consider the explicit enforcement of orthogonality as a special case resting within a more general framework, rather than an ad-hoc approach.

The approach does not always produce the results we wish for. Most obviously it equivocates between the front and back views, say; our current solution is to append directly opposite views. One possible resolution (if desired) would be further processing, front views often exhibit greater variation than back views. Another might be to better model the Mahalanobis surface, perhaps using a Gaussian Mixture Model to determine a surface of equal likelihood. This might resolve those few cases where even the explicit enforcement of orthogonality gives incorrect results (see the doll on the bottom row of Figure 3).

The method works for some but not all asymmetric objects. The violin model in Figure 6 shows a positive result, yet even in cases of failure the general shape of the Mahalanobis surface remains intact. This suggests that it may pay to conduct future analysis continue along the lines begun here.

A deeper comparison with the work of others may benefit this study, but our method is undoubtedly more simple and more efficient; it is more versatile in the sense it supports wider definitions of canonical view; testing robustness should be the objective of future experiments. It might be interesting to map onto the Mahalanobis surface canonical views that humans have chosen as canonical.

We conclude that our method reliably produces canonical views in a simple way, and that the central role the Mahalanobis surface plays deserves further study.

References

[1] V. Blanz, M.J. Tarr, H.H. Bülthoff, and T. Vetter. What object attributes determine canonical views? Technical Report 42, Max–Plank–Institut für bioligische Kybernetik, 1996.

[2] T. Denton, M.F. Demirci, J. Abrahamson, A. Shokoufandeh, and S. Dickson. Selecting canonical views for view-based 3-d object recognition. In *Proc. International Conference on Pattern Recognition*, pages 23–26, 2004.

[3] W.T. Freeman. The generic viewpoint assumption in a framework for visual perception. *Nature*, 368(6471):542–545, 1994.

[4] P. Hall, D. Marshall, and R. Martin. Merging and splitting eigenspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):1042–1049, September 2000.

[5] R. Harvey, A. Bosson, and J.A. Bangham. The robustness of some scale-spaces. In *British Machine Vision Conference*, pages 11-20, August 1997.

[6] W. Hong, A.Y. Yang, K. Huang, and Y. Ma. On symmetry and muliple view geometry: Structure, pose, and calibration from a single image. *International Journal of Computer Vision*, 60(3):241–265, 2004.

[7] S. Palmer, R. Rosch, and P. Chase. Canonical perspective and the perception of objects. In *Attention and Performance IX*, pages 135–151, 1981.

[8] D.I. Perrett and M.H. Harries. Characteristic views and the visual inspection of simple faceted and smooth objects: ”tetrahedra and potatoes”. *Perception*, 17:703–720, 1988.

[9] G. Peters, B. Zitova, and C. von der Malsburg. How to measure the pose robustness of object views. *Image and Vision Computing*, 20:341–348, 2002.

[10] S. Obdržálek and J. Matas. Object recognition using local affine frames on distinguished regions. In *Proc. 13th British Machine Vision Conference*, pages 113–132, 2002.

[11] D. Weinshall and M. Werman. On view likelihood and stability. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):97–108, 1997.

[12] P. Min, J. Chen, and T. Funkhouser. A 2D Sketch Interface for a 3D Model Search Engine. *SIGGRAPH 2002 Technical Sketch*, July 2002.