Automatic Selection of Reference Views for Image-Based Scene Representations*

Václav Hlaváč[1], Aleš Leonardis[2], Tomáš Werner[1]

[1] Czech Technical University, Faculty of Electrical Engineering
Department of Control Engineering
12135 Prague 2, Karlovo náměstí 13, Czech Republic
{hlavac,werner}@vision.felk.cvut.cz
[2] Technical University Vienna
Department of Pattern Recognition and Image Processing
A-1040 Wien, Treitlstrasse 3, Austria
(also at the University of Ljubljana, Slovenia)
ales@prip.tuwien.ac.at

Abstract. The problem addressed in this paper is related to displaying a real 3-D scene from any viewpoint. To display a scene, a relatively sparse set of 2-D *reference views* is stored. The images that are in-between the reference views are obtained by *interpolation* of coordinates and brightness (colour). This approach is able to generate the scene representation and render images automatically and efficiently even for complex scenes of 3-D objects. This is possible since the processing time does not depend on the complexity of the scene as there is no attempt to understand the semantics of images.

In this paper we present *a novel approach* to automatically determine a minimal set of views from which the complete scene can be rendered. The method consists of two procedures: view-interval growing and selection. The first procedure independently searches for the intervals from which large portions of the scene can be rendered. These intervals are then passed to the selection procedure, which selects the minimal set of necessary views. The selection procedure is posed as an optimization problem that minimizes the number of reference views and the error due to the interpolation.

* We are primarily indebted to Prof. Roger Hersch from the EPFL Lausanne in Switzerland who initiated the whole project and mediated T. Werner's support from the Swiss National Fund, grant 83H-036863. We were supported by the Grant Agency of the Czech Republic, grants 102/93/0954, 102/95/1378, and the EU grant Copernicus No. 1068. A. Leonardis' work was supported partly by the Austrian National Fonds zur Förderung der wissenschaftlichen Forschung under grant S7002MAT, the Ministry of Science and Technology of Republic of Slovenia (Project J2-6187), and by U.S.-Slovene Joint Board (Project #95-158). We are grateful to Tomáš Pajdla, Walter Kropatsch, and Radim Šára for discussions.

1 Task Formulation

Traditional approaches for displaying a 3-D scene from any viewpoint use a full 3-D model of the scene. The drawback is that the methods for automatic 3-D model reconstruction can only cope with simple man-made objects.

Recently, alternative methods have appeared using *image-based scene representations*. In this case, the scene is not represented by a 3-D model but rather by a collection of 2-D views. The research activities focus either on the geometry of the problem [UB91, Sha94, LF94], or on the image interpolation problem [CW93, SD95, WHH95]. However, none of the works address the problem of how to determine an optimal set of views from which the scene can be rendered to a sufficient degree of accuracy.

This paper proposes a *novel paradigm* of how to automatically select the set of reference views from the set of all views of a scene for displaying purposes. In further text, the term *primary views* denotes captured 2-D images that constitute the input information about the scene. By *reference views* we understand the smallest subset of the original set of primary views that is sufficient to accurately interpolate all other primary views.

Let us illustrate the above idea on an example. Imagine, that a 3-D rigid object is fixed on a table that can turn and slant. A camera looks at the object on the table from a fixed point. We can capture and store 2-D views covering the whole hemisphere of possible views. Assume for the moment that we already have a set of reference views. The reference views should encompass the crucial information needed for displaying. Other views can be computed by interpolation from the reference views, where the interpolation combines both position (coordinates) and brightness or colour. These images are called *interpolated views*.

There are four issues related to the formulated task [WHH95]: (1) How to predict the position and the intensity of a point in the new view if the positions and the intensities of corresponding points in the reference views are known? (2) How to determine the visibility of points in the new view? (3) How to find the optimal set of necessary reference views? (4) How to find the correspondence between reference views? In this paper we primarily concentrate on *issue 3.*

A "brute force" method would represent a scene by generating a densely sampled set of primary views. Our intention is to do better and replace many primary views by a much sparser set of reference views. Practically, we can think of several possible setups related to the capturing of primary views: (1) A 3-D object can be placed on a turn-and-slant table and captured by a camera from a fixed viewpoint. (2) Several dozens of cameras can be placed around the captured object. This setup was proposed for the so called virtualized reality [KNR95]. (3) The source of primary views can also be a video sequence. An uncalibrated camera moving along an unknown trajectory is the most general case.

Note, that the practical experiments reported in this paper fall in the first category, however, our paradigm applies to all three categories.

2 New Paradigm for Selection of Reference Views

The *basic idea* is to group close and similar primary views and represent them by the reference views. Let us first assume the simple case where the object makes a 1-DOF movement with respect to the camera. This corresponds to an object placed on a turn-table and observed by a camera from a fixed distance. Moreover, let all possible primary views lie on a view circle. Let us assume that a densely sampled set of primary views is captured. Similar and neighbouring views on the view circle can be used to create the *viewing interval* of the view angle. Limits of the interval correspond to two 2-D images. The images within the interval can be interpolated from the extremal 2-D images. If there is a close similarity between the interpolated and the primary views, we can represent the whole viewing interval by its two extremal views. They will be called reference views. The similarity between interpolated and primary views is based on an approximately identical visual appearance. The *visual appearance dissimilarity* (VAD) measure is used for comparing interpolated and primary views. The VAD is described in more detail in Section 2.1. The VAD should be below a predefined precision for all views in the interval.

Since we do not want to be affected by the starting point during the growth of the viewing interval, we start the process of creating the interval from each primary view independently. The result is a set of candidate intervals that can potentially overlap. To select the best set that covers the whole range of views is the goal of an optimization task.

In this paper, the stress is put on the method of computing reference views. We worked out the simplest case where primary views lie on the view circle. The generalization to the view sphere case seems possible. A few remarks on the generalization will be discussed in Section 4.

2.1 The Visual Appearance Dissimilarity

The *visual appearance dissimilarity* (VAD) assesses the closeness of the match between a primary view and its interpolation created from some reference views. In other words, it expresses the error due to replacing the real image by the interpolated one. The word *dissimilarity* can hardly be defined in absolute terms. It is related to the purpose the interpolated images will serve to and must be designed with this purpose in mind. For instance, a human observer can tolerate slight geometrical distortion of the whole image, but is very disturbed by some spatially small artifacts in the images. For the demonstration of our approach, we chose the following simple VAD function.

Let $v(x,y)$ denote the primary view and $i(x,y)$ the interpolated view. Let $\mathcal{O}_i$ denote the set of points that could not be interpolated in $i(x,y)$ because of occlusion. A complementary set $\mathcal{O}_i^C$ denotes the points that are not occluded. We can define a VAD as a scalar function that is basically independent on image interpretation:

$$\mathrm{VAD}(v,i) = \alpha\ |\mathcal{O}_i| + \beta \sum_{(x,y)\in\mathcal{O}_i^C} |v(x,y) - i(x,y)|\ , \qquad (1)$$

where α and β weight the influence of occlusion and of the error due to incorrect matching or interpolation, respectively. $|.|$ is the number of elements of a set. Another possibility is to use the sum of squared pixelwise differences or other measures [Cha89].

2.2 Constructing the Set of Plausible Viewing Intervals

Let us consider the simple 1-DOF case where primary views lie on the view circle. Then the ***plausible viewing interval*** (PVI) is a viewing interval such that all primary views inside it can be interpolated from the two primary views corresponding to PVIs endpoints with the VAD smaller than some predefined value ϵ.

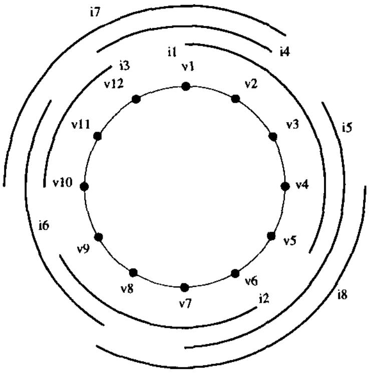


Fig. 1. Example of primary views and PVIs.

Let us focus now on the construction of the set of PVIs. The input is a set of all primary views, $v_1, \ldots, v_N$. The neighbour of the view v_N is the view v_1. The desired output is the set of the PVIs. In Fig. 1, the dots on the circle represent primary views, and the circular arcs illustrate constructed PVIs, $i_1, \ldots, i_8$.

We use a technique similar to region growing. Each PVI has a starting and an ending primary view. Initially, we have a degenerate case where each PVI consists of only one primary view. Then we can try to extend the PVI by adding the left neighbouring primary view to the existing PVI. If the expanded interval is to be the PVI, we should check if all interpolated intermediate views and the corresponding primary views satisfy the condition[3] that the image VAD $< \epsilon$. If the left neighbour fails, we try to merge the right neighbour. If the right neighbour also fails, then the final PVI is obtained.

The interval merging process results in one PVI for each primary view. Now, the selection procedure, described in section 2.3, has the task to find those PVIs that best cover the whole view circle.

All the intervals are first grown to their full extent and then passed to the selection module. As a consequence of the selection process, eventually very few of the intervals emerge as sufficient to cover the view-space. We call this strategy *Build(intervals)-then-Select* because it grows all the intervals fully and independently, and then discards the redundant ones. However, the computational cost can be reduced. Instead of growing all the intervals completely, it is possible to discard some of the redundant intervals even before they are fully grown [Leo93]. This suggests incorporating the selection procedure into the procedure of growing intervals. We call this approach *Build(intervals)-and-Select*.

[3] In current experiments, we check the VAD for all intermediate images. However, repeating the VAD check for all primary views within the PVI every time the interval is extended is time consuming. If the current PVI was interpolable from extremal images, it is likely that, after merging one new primary view to it, we might not check every primary view for the $VAD < \epsilon$ again.

2.3 Selection of the Optimal Set of Plausible Viewing Intervals

The selection is a discrete optimization problem and its formulation is similar to the one presented in [Leo93].

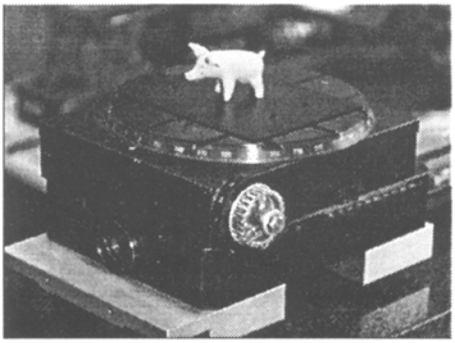


Fig. 2. The experimental setup for capturing the sequence.

The set of PVIs which has been generated may be highly redundant. Thus the selection procedure selects a subset of best PVIs and rejects the superfluous ones. The optimization of an objective function including the information about the competing PVIs is carried out. The objective function is proposed in the following form:

$$F(\mathbf{i}) = \mathbf{i}^T\mathbf{C}\mathbf{i} \qquad (2)$$

The vector $\mathbf{i}^T = [i_1, i_2, \ldots, i_R]$ denotes a set of PVIs, where i_i is a *presence-variable* having the value 1 for the presence and 0 for the absence of the i-th PVI in the resulting set of PVIs. The diagonal terms of the matrix **C** express the cost-benefit value for a particular PVI

$$c_{ii} = \mathrm{K}_1 s_i - \mathrm{K}_2||\epsilon_i|| - \mathrm{K}_3 V_i \ , \qquad (3)$$

where V_i is the number of reference views plus the data volume needed for correspondence per PVI, s_i is the number of primary views that are covered by the interval, and $||\epsilon_i||$ is the error measure calculated for the interval. The coefficients K_1, K_2, and K_3 adjust the contribution of the three terms.

The off-diagonal terms handle the interaction between the overlapping PVIs.

$$c_{ij} = \frac{-\mathrm{K}_1|D_i \cap D_j| + \mathrm{K}_2||\epsilon_{ij}|| + K_3\theta_{ij}}{2} \ , \quad ||\epsilon_{ij}|| = \max(\sum_{D_i \cap D_j} \epsilon_i, \sum_{D_i \cap D_j} \epsilon_j) \qquad (4)$$

D_i denotes the domain of the i-th PVI, i.e., the primary views that are covered by the i-th PVI. $\theta_{ij} = 1$ if the intervals i and j are adjacent, i.e., if the end view of the first interval is equal to the start view of the other interval or vice versa. Otherwise, $\theta_{ij} = 0$.

The objective function takes into account the overlap between the different PVIs which may be completely or partially overlapped. However, we consider only the pairwise overlaps in the final solution. The matrix **C** is symmetric, and depending on the overlap, it can be sparse or banded. These properties of matrix **C** can be used to reduce the computations needed to calculate the value of $F(\mathbf{i})$.

We have now formulated the selection problem in such a way that its solution corresponds to the global extreme of the objective function. Maximization of the objective function $F(\mathbf{i})$ belongs to the class of combinatorial optimization problems (quadratic Boolean problem). Since the number of possible

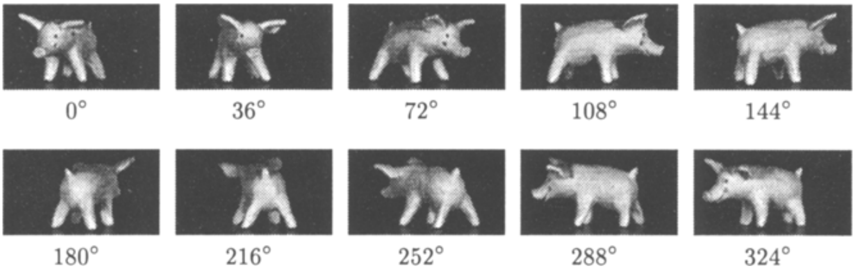


Fig. 3. Ten views from the PIG sequence 36 degrees apart.

solutions increases exponentially with the size of the problem, the exhaustive search is usually not tractable. Various methods have been proposed for finding a global "optimum" of a class of nonlinear objective functions. Among these methods are winner-takes-all strategy, simulated annealing, microcanonical annealing, mean field annealing, Hopfield networks, continuation methods, and genetic algorithms. We are currently using two different optimization methods. One is a simple greedy algorithm that is computationally very efficient, the other one is tabu search [GL93, SL95]. Tabu search is computationally a little more demanding but it provides consistently better results than the greedy algorithm.

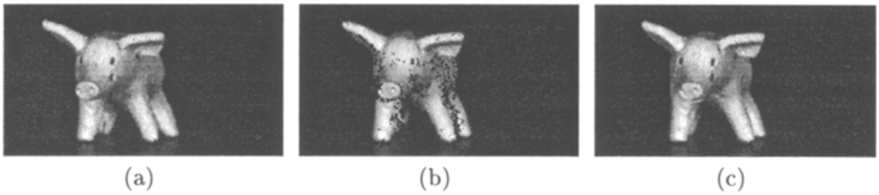


Fig. 4. A pair of reference views and the interpolated view. (a, c) are the reference views (8° and 16° in the captured sequence), (b) is the interpolated view (corresponding to 12° in the sequence).

3 Experiments

In this section, we show the practical feasibility of our approach on an experiment which was carried out on real data and for one degree of freedom of the object's motion. The *Build(intervals)-then-Select* control strategy was used to select the optimal set of reference views from a large set of primary views.

The setup which was used to capture primary views is shown in Fig. 2. The object to be captured, a small clay sculpture of a pig about 6 cm long, was placed on a turn-table. The object was viewed by the camera from the distance 2 metres (the focal length of the camera lens was 100 mm, CCD chip had the side 12.7 mm). That means that the projection was close to orthographic.

We captured 180 primary views of the object. Each neighbouring pair of views was 2 degrees apart. Ten primary views of the sequence are shown in Fig. 3. The size of the images was 150 x 80 pixels.

Assuming for a moment that a set of reference views of these 180 primary views have already been selected, the remaining views can be constructed by view interpolation as described, e.g., in [LF94, WHH95]. The example triplet of two reference and one interpolated views is shown in Fig. 4. In the interpolated view, the pixels detected by the stereo matcher as occluded can be seen as black pixels inside the boundary of the object. In these pixels the view interpolation fails.

The view interpolation requires that the correspondence is available for each pair of neighbouring reference views. For that, we used the binocular stereo matcher [CHMR92]. This algorithm provides correspondence as well as occlusion maps. The algorithm requires that the epipolar lines be parallel to scanlines in the matched image pair, which is ensured by our experimental setup (the projection is close to orthographic) without additional rectification.

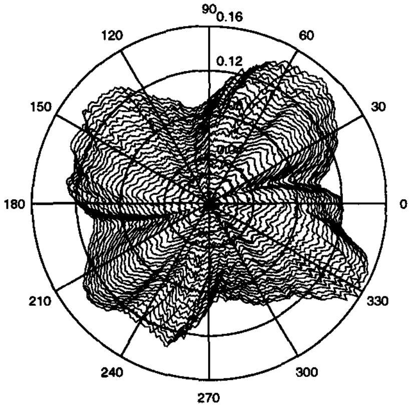


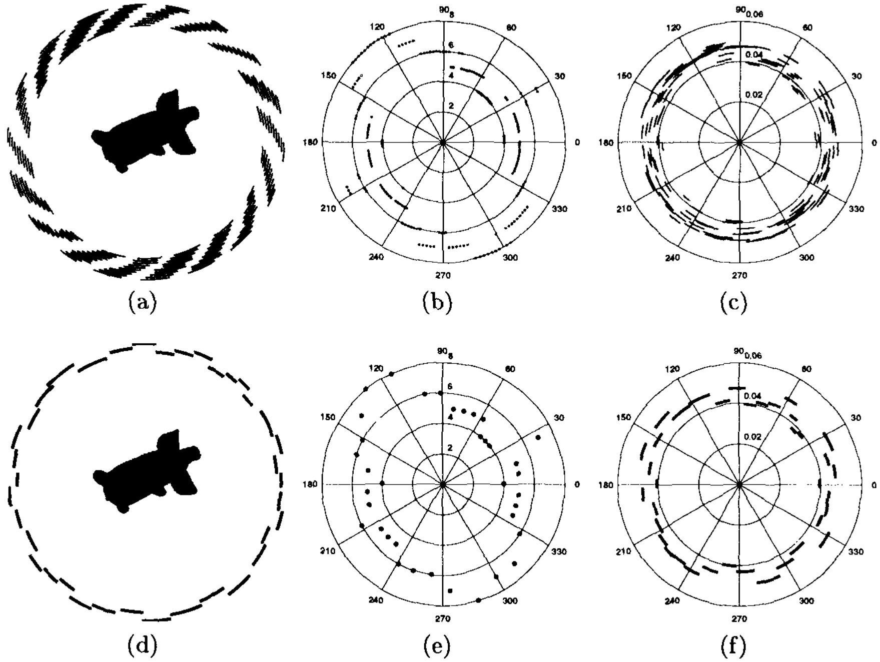
Fig. 5. The plot of VAD defined by Eq. (5) for the PIG sequence.

For the VAD, we used only the first term of Eq. (1). The reason for this simplification is that the interpolated views have always slight geometrical distortion in our current implementation, which degrades the result of the direct pixelwise image comparison (as in Eq. (1)). Thus we used

$$\mathrm{VAD}(v,i) = \frac{|\mathcal{O}_i|}{|\mathcal{N}_i|}, \tag{5}$$

where $\mathcal{O}_i$ is the set of pixels in the interpolated image i that cannot be computed due to occlusion, and $\mathcal{N}_i$ is the set of all pixels in the interpolated image i. This function is very simple, e.g., the primary view v does not occur explicitly in it. In most cases, the function is almost constant while i moves from the first reference view to the second one. The VAD (more accurately, its mean value over the interval between the pair of reference views) for our object is plotted in Fig. 5. Here, the interpolated views were constructed from a pair of reference views using correspondence obtained by the stereo matcher [CHMR92]. The angle in the plot in polar coordinates represents the angle of the first view of the reference view pair. Each curve was obtained for different angular distance between the two reference views (the most inner curve corresponds to the distance 1°, the most outer one to 30°, the step is 1°). The radius represents the mean value of the VAD over the interpolated interval.

PVIs were built, starting independently from each primary view (see section 2.2). The merging process was limited by the condition VAD < 0.05. Then, in the selection procedure, the optimal set of PVIs was found as it is described in the section 2.3. The parameters in the objective function (Eq. (3) and (4)) were chosen $V_i = 3$ (the data volume needed for storing the correspondence was approximately equal to the data volume needed for storing one image), $K_1 = 1$, $K_2 = 1$, and $K_3 = 1$. We observed that the results remain the same even for large variations of the parameters. To find a solution of Eq. (2), the tabu search was used.



(a) (b) (c)

(d) (e) (f)

Fig. 6. The results of building PVIs (a, b, c) and selecting the optimal set of the built PVIs (d, e, f).

The result of the process is shown in Fig. 6. Subfigures (a, b, c) show the result of the growing process, subfigures (d, e, f) show the results of the selection process. The intervals are depicted as arcs in (a, d). The radius coordinate has no other meaning here than to help visualizing the overlapping intervals. The top view of the object is superimposed. The length of the PVIs is plotted in (b, e). The intervals are depicted as dots. The radius represents the numbers of views the PVI covers. The VAD for each PVI is plotted in (c, f).

4 Generalization to the Case of the Whole View Sphere

The automatic selection of reference views for image-based scene representations has been demonstrated for the 1-DOF viewing directions (view circle). There is no principal obstacle to generalize the paradigm to more difficult 2-DOF case, where possible viewpoints lie on a view sphere. The selection procedure remains the same. The only difference is in creating plausible view patches (analogous to current PVIs). The plausible view patches can be created from primary views independently using similar approaches like in region-growing based segmentation.

The growing process becomes very computationally intensive. Therefore, more attention should be paid to utilizing *Build(intervals)-and-Select* control strategy, where the selection procedure is started already before the patches are fully grown. This rules out apparently redundant regions before they are fully grown.

5 Conclusions and Future Work

The contribution of this paper is the paradigm formulation that allows to automatically obtain the optimal set of 2-D reference views. This is relevant to currently frequently studied image-based scene representations.

Experiments on real data have demonstrated the result of using our algorithm for a set of 180 primary views. Yet how should the quality of the results be evaluated? The problem is that this evaluation is closely related to VAD. Although we are aware that our current VAD is probably inadequate for real requirements and thus opened for further research, the main idea of this paper can be separated from the specific dissimilarity measure.

In the future, we would like to dedicate more attention to VAD. The difficulty is the discrepancy between the expected visual appearance of the interpolated views (complicated and not well understood process studied in cognitive sciences) and extremely simplified dissimilarity function used for comparing interpolated and primary views. If we had better VAD, we would just incorporate it in the current framework. More extensive quantitative comparisons will be necessary as well as further improvements of geometrical validity of interpolated views. In fact, the latter is another separate track of our current research.

The extension to 2-DOF (view hemisphere) is the step we plan to do next. Using the *Build(intervals)-and-Select* control strategy, we would like to lower the computational demands so that the 2-DOF case would become practical.

References

[CW93] S.E. Chen and L. Williams. View interpolation for image synthesis. In *Proceedings of the SIGRAPH'93*, pages 279–288, 1993.

[GL93] F. Glover and M. Laguna. Tabu search. In C. R. Reeves, editor, *Modern heuristic techniques for combinatorial problems*, pages 70–150. Blackwell Scientific Publications, 1993.

[Cha89] Shi-Kuo Chang. *Principles of Pictoral Information Systems Design.* Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1989.

[CHMR92] Ingemar J. Cox, Sunita Hingorani, Bruce M. Maggs, and Satish B. Rao. Stereo without disparity gradient smoothing: a Bayesian sensor fusion solution. In *British Machine Vision Conference*, pages 337–346, Berlin, 1992. Springer-Verlag.

[KNR95] T. Kanade, P.J. Narayanan, and P.W. Rander. Virtualized reality: Concepts and early results. In *Proceedings of the Visual Scene Representation Workshop, Boston, MA., USA, June 24*, page 8. available via http://www.cis.upenn.edu/ẽero/VisSceneRep95.html, 1995.

[Leo93] A. Leonardis. *Image Analysis Using Parametric Models.* PhD thesis, University of Ljubljana, Slovenia, March 1993.

[LF94] Stéphane Laveau and Olivier Faugeras. 3-D scene representation as a collection of images. In *Proc. of 12th International Conf. on Pattern Recognition, Jerusalem, Israel*, pages 689–691, October 9–13 1994. Also the technical report, available on ftp.inria.fr.

[SD95] S.M. Seitz and C.R. Dyer. Physically-valid view synthesis by image interpolation. In *Proceedings of the Visual Scene Representation Workshop, Boston, MA., USA, June 24*, page 8. available via http://www.cis.upenn.edu/ẽero/VisSceneRep95.html, 1995.

[Sha94] A. Shashua. Trilinearity in visual recognition by alignment. In *Proceedings of the 3rd European Conference on Computer Vision, Stockholm, Sweden*, volume 1, pages 479–484. SV, 1994.

[SL95] M. Stricker and A. Leonardis. ExSel++: A general framework to extract parametric models. In V. Hlaváč and R. Šára, editors, *6th CAIP'95*, number 970 in Lecture Notes in Computer Science, pages 90–97, Prague, Czech Republic, September 1995. Springer.

[UB91] S. Ullman and R. Basri. Recognition by linear combination of models. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 13(10):992–1005, October 1991.

[WHH95] T. Werner, R.D. Hersch, and V. Hlaváč. Rendering real-world objects using view interpolation. In *ICCV95*, pages 957–962, Boston, USA, June 1995. IEEE Press.