A Systematic Approach on 3D Object Modeling and Recognition

Laxmi Devi
MCA, M.Phill, Department of Computer Science, Global Open University, Nagaland, India

ABSTRACT

This article introduces a novel representation for three-dimensional (3D) objects in terms of local affine-invariant descriptors of their images and the spatial relationships between the corresponding surface patches. Geometric constraints associated with different views of the same patches under affine projection are combined with a normalized representation of their appearance to guide matching and reconstruction, allowing the acquisition of true 3D affine and Euclidean models from multiple unregistered images, as well as their recognition in photographs taken from arbitrary viewpoints. The proposed approach does not require a separate segmentation stage, and it is applicable to highly cluttered scenes. Modeling and recognition results are presented.

Keywords: Three-dimensional object recognition, image-based modeling, affine-invariant image descriptors, multiview geometry.

INTRODUCTION

This article addresses the problem of recognizing three-dimensional (3D) objects in photographs. Traditional feature-based geometric approaches to this problem— such as alignment or geometric hashing enumerate various subsets of geometric image features before using pose consistency constraints to confirm or discard competing match hypotheses, but they largely ignore the rich source of information contained in the image brightness and/or color pattern, and thus typically lack an effective mechanism for selecting promising matches. Appearance-based methods—as originally proposed in the context of face recognition and 3D object recognition take the opposite view, and prefer to explicit geometric reasoning a classical pattern recognition framework that exploits the discriminatory power of (relatively) low-dimensional, empirical models of global object appearance in classification tasks. However, they typically deemphasize the combinatorial aspects of the search involved in any matching task, which limits their ability to handle occlusion and clutter.

Concretely, we propose using local image descriptors that are invariant under affine transformations of the spatial domain and of the brightness/color signal to capture the appearance of salient surface patches, and a set of multi-view geometric constraints related to those studied in the structure from motion literature to capture their spatial relationship. Our approach is directly related to a number of recent techniques that combine local models of image appearance in the neighborhood of salient features— or “interest points” (Harris and Stephens, 1988)—with local and/or global geometric constraints in wide-baseline stereo matching image retrieval and object recognition tasks. These methods normally either require storing a large number of views for each object or limiting the range of admissible viewpoints. In contrast, our approach supports the automatic acquisition of explicit 3D affine and Euclidean object models from multiple unregistered images, and their recognition in heavily-cluttered pictures taken from arbitrary viewpoints.

The rest of this presentation is organized as follows: Section 2 presents the main elements of our approach. Its applications to 3D object modeling and recognition are discussed in Sections 3 and 4. In practice, object models are constructed in controlled situations with little or no clutter, and the stronger consistency constraints associated with 3D models make up for the presence of significant clutter and occlusion in recognition tasks, avoiding the need for a separate segmentation stage. Modeling and recognition examples can be found in Figures 1, 14–15, 19 and 25, and a detailed description of our experiments, including quantitative recognition results. We conclude in Section 5 with a brief discussion of the promise and limitations of the proposed approach.

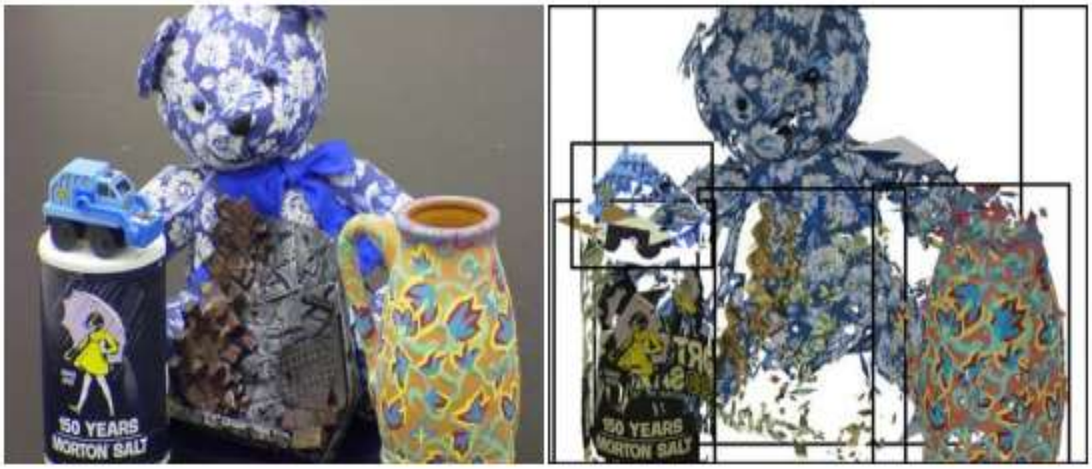

Figure 1. Results of a recognition experiment. Left: A test image. Right: Instances of five models (a teddy bear, a doll stand, a salt can, a toy truck and a vase) have been recognized, and the models are rendered in the poses estimated by our program. Bounding boxes for the reprojections are shown as black rectangles.

SYSTEMATIC APPROACH

Affine Regions and their Description

The construction of local invariant models of object appearance involves two steps, the detection of salient image regions, and their description. Ideally, the regions found in two images of the same object should be the projections of the same surface patches. Therefore, they must be covariant, with regions detected in the first picture mapping onto those found in the second one via the geometric and photometric transformations induced by the corresponding viewpoint and illumination changes. In turn, detection must be followed by a description stage that constructs a region representation invariant under these changes. For small patches of smooth Lambertian surfaces, the transformations are (to first order) affine, and we use the approach recently proposed by Mikolajczyk and Schmid to find the corresponding affine regions: Briefly, the algorithm iterates over steps where (1) an elliptical image region is deformed to maximize the isotropy of the corresponding brightness pattern (shape adaptation); (2) its characteristic scale is determined as a local extremum of the normalized Laplacian in scale space (scale selection); and (3) the Harris operator is used to refine the position of the the ellipse's center (localization). The scale-invariant interest point detector proposed in provides an initial guess for this procedure, and the elliptical region obtained at convergence can be shown to be covariant under affine transformations. The affine region detection process used in this chapter implements both this algorithm and a variant where a difference-of-Gaussians (DoG) operator replaces the Harris interest point detector. Note that this operator tends to find corners and points where significant intensity changes occur, while the DoG detector is (in general) attracted to the centers of roughly uniform regions (blobs): Intuitively, the two operators provide complementary kinds of information as in Figure 2.

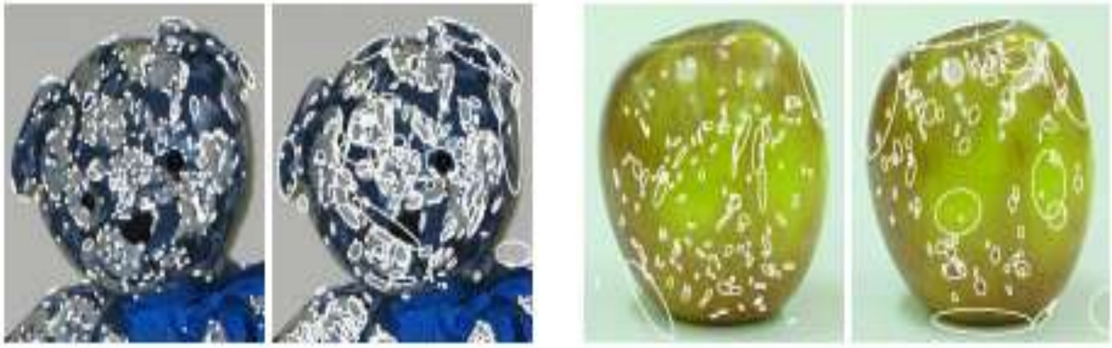

Fig. 2. Affine regions found by Harris-Laplacian (left) and DoG (right) detectors.

The affine regions output by our detection process are ellipses that can be mapped onto a unit circle centered at the origin using a one-parameter family of affine transformations. This ambiguity can be resolved by determining the dominant gradient orientation of the image region, turning the corresponding ellipse into a parallelogram and the unit circle into a square. Thus, the output of the detection process is a set of image regions in the shape of parallelograms, together with affine rectifying transformations that map each parallelogram onto a “unit” square centered at the origin (Figure 3).

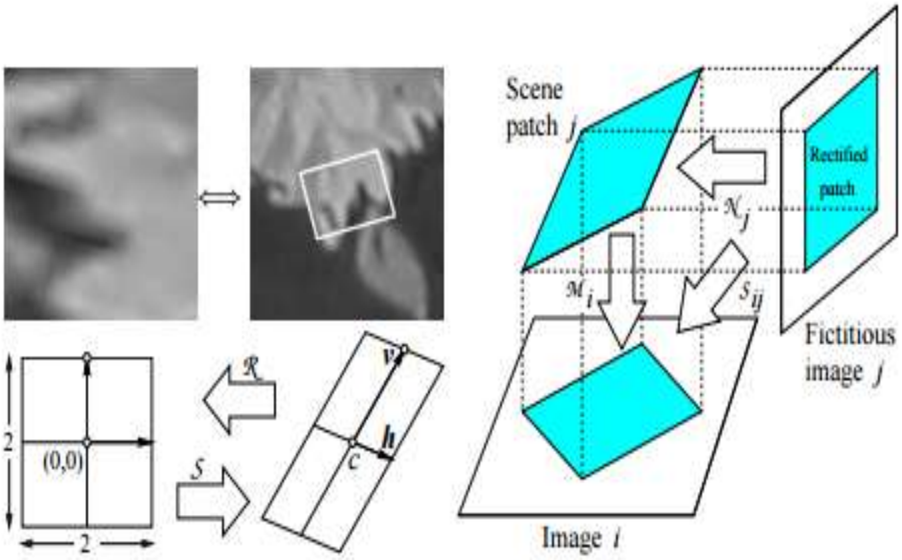

Fig. 3. Geometric structure. Top left: A rectified patch and the original image region. Bottom left: Interpretation of the rectification matrix R and its inverse S. Right: Interpretation of the decomposition of the mapping Sij into the product of a projection matrix Mi and an inverse projection matrix Nj

ANALYSIS ON 3D MORPHABLE MODELS

Morphing between 3D objects is a well-known computer graphics technique. 3D morphable face models apply the general concept into the vector space representation of face models. The main idea behind the morphable face model approach is that given a sufficiently large database of 3D face models any arbitrary face can be generated by morphing between the ones in the database. The 3D morphable model used in this thesis was developed by Volker Blanz and Thomas Vetter [3], who extended the 2D approach in. The generation of 3D head models was done in collaboration with their lab. Their database of 3D models was built by recording the faces of 200 subjects with a 3D laser scanner. Then 3D correspondences between the head models were established in a semi-automatic way using techniques derived from optical flow computation. Using these correspondences, a new 3D face model can be generated by morphing the existing models in the database.

To create a 3D face model from a set of 2D face images, an analysis by synthesis loop is used to find the morphing parameters such that the rendered images of the 3D model are as close as possible to the input images. These parameters include shape and texture coefficients, illumination, orientation, and face position. The optimization algorithm starts with manual alignment of the average face (of the 200 head models) with the face in the image. Iteratively, the algorithm attempts to minimize the error between the synthetic reconstruction at that point with the input image with respect to the the sum of square errors over all color channels and all pixels.

Training Images Generated from Face Models Morphable models allow for the simplistic generation of 3D face models which are used in the training of the face recognition system. First, high quality frontal and halfprofile pictures are taken of each subject under ambient lighting conditions. These images are then used used as input to the analysis by synthesis loop which yields a face model. This face model can be used to graphically render synthetic face images under varying pose and illumination conditions. Examples of the pairs of input images and corresponding synthetic images created by rendering the 3D face models are shown in Figure 4.

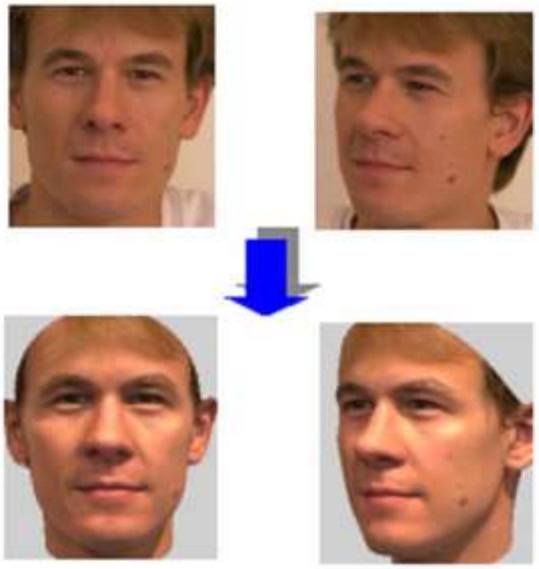

Figure 4: Generation of the 3D model. The top images are the real images used to generate a 3D model. The bottom images are synthetic images generated from the model. Notice the similarity between the original and synthetic images.

Be created by rendering the models. The 3D morph able model also provides the full 3D correspondence information which allows for automatic extraction of facial components and facial regions. This greatly simplifies the problem of obtaining the large quantity of training images required to train a robust face recognition system. The 3D face models were used in the creation of the positive training set. The negative training set was built from randomly extracting 13655 patterns of size 58×58 from a database of non-face images. Synthetic faces were generated at a resolution of 58×58 for the 6 subjects under varying pose and illumination conditions. Specifically, the faces were rotated in depth from −34◦ to 34◦ in 2◦ increments. Two illumination models were used to simulate real lighting conditions. One model consisted of ambient light alone, while the other model was composed of directed light in addition to ambient light. The directed light was pointed at the center of the face and positioned between −90◦ and +90◦ in azimuth and 0◦ and 75◦ in elevation. The angular position of directed light was incremented by 15◦ in both directions. This training set was used to train both the component-based and global detection units.

3D OBJECT RECOGNITION

We now assume that the modeling approach presented in Section 3 has been used to create a library of 3D object models, and address the problem of identifying instances of these models in a test image. In many respects, this process is analogous to the method described in below Section for pairwise image matching. As before, Algorithm 1 outlines the overall process. The parameters used for Algorithm 1 in this setting are given in. Further details are given in the rest of this section.

RANSAC-LIKE SELECTION/ESTIMATION PROCEDURE

As noted in Section 2, various methods for finding matching features consistent with a given set of geometric constraints have been proposed in the past, including interpretation tree—or alignment—techniques, and robust statistical methods such as RANSAC and its variants. Both alignment and RANSAC can easily be implemented in the context of Algorithm 1. We have experimented with several alternatives: The first one is a recursive implementation of alignment where an interpretation tree is visited in a depth-first manner (null matches between model patches and “empty” image regions being used to handle occlusion and faulty detection) until a maximum depth N is reached (N = 20 in our experiments), or the mean reprojection error exceeds E in all branches up to that depth.

We have also implemented plain RANSAC and two variants: a “greedy” version where, as before, M groups of matches of size lesser than or equal to N are chosen in a deterministic, greedy manner to minimize the mean projection error, and used instead of random samples; and an "exhaustive" version where all pairs of candidate matches are examined. The computational costs of the RANSAC variants are easy to estimate. The cost of alignment is more difficult to assess, but can be shown to be a low-order polynomial in the size n of the model when there is little or no clutter, and exponential in n in the presence of clutter when no limit on the depth of the tree search is imposed (Grimson, 1990). The worst-case computational complexity of our bounded tree search is O(nN), but determining its expected cost is beyond the scope of this paper. As will be shown in Section 4.5, the "greedy" version of RANSAC has performed best in our experiments.

EXPERIMENTAL ANALYSIS

Our recognition experiments match all eight of our object models against a set of 30 images (the photograph from and the 30 pictures shown in Figure 5). Each image contains instances of up to five object models, even though most of them only contain one or two. gives quantitative recognition results for the different "black-and-white" variants of our algorithm, where color information is not used. The parameters for these tests are fixed to their nominal values of m = 10, a = 0.1, and d = 0.15. With these settings, none of the methods tested gives false positives, and the "greedy" version of RANSAC with N = 20 gives the best performance, with a recognition rate (averaged over the eight object models) of 88%. The time costs as given in the table are per image-object combination, in minutes. Since it has consistently performed best in our experiments, we will from now on focus on the greedy variant of RANSAC with N = 20. It is interesting to compare different image descriptors and to test whether the use of color information may boost recognition performance. Shows the results of a quantitative experiment: It can be seen that the combination of color and SIFT gives the best performance, with a mean recognition rate of 94%. (This rate is for the nominal settings of the detection parameters. The effect of these parameters is discussed below.)

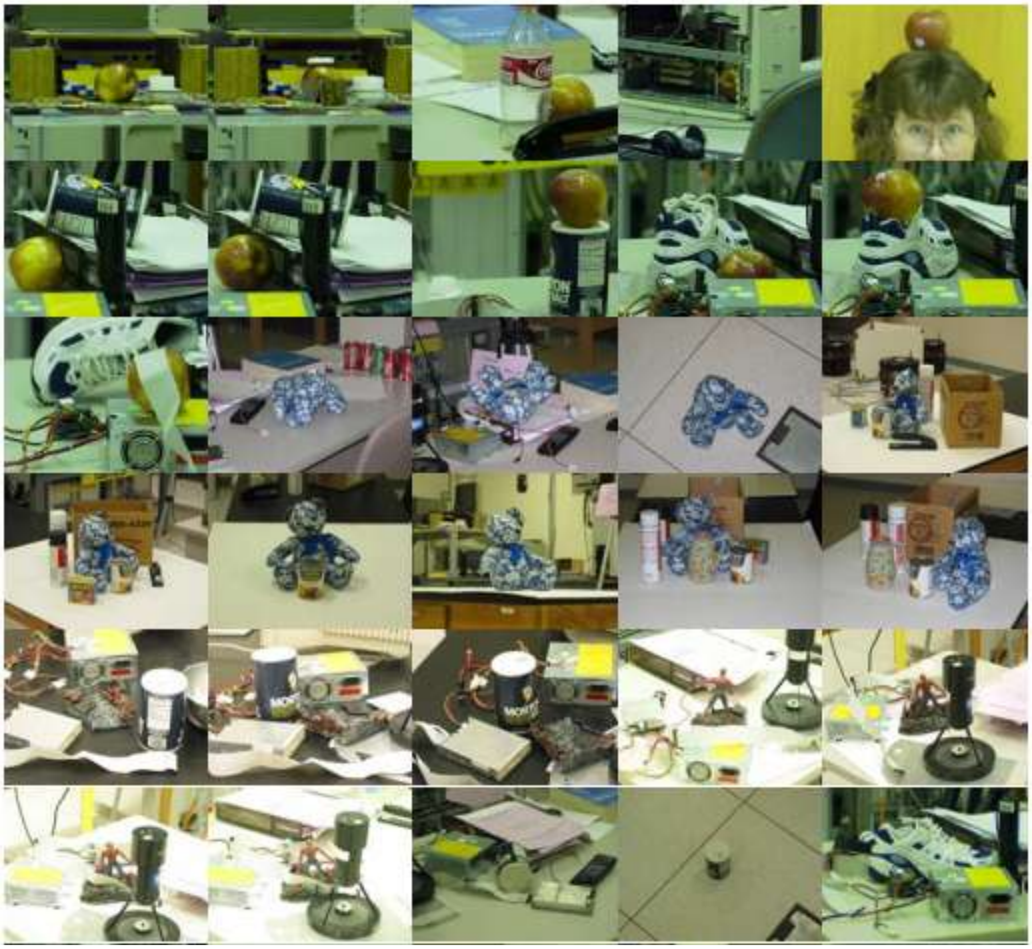

Figure 5. The dataset used in our recognition experiments: 30 of the images are shown here.

CONCLUSIONS

This paper involved a new development face recognition with the incorporation of 3D morphable models and component-based face recognition. This combination allowed the training of a face recognition system which required only two face images of each person. From these two images, 3D face models were computed and then used to render a large number of synthetic images under varying poses and lighting conditions. These synthetic images were then used to train a component-based face detection and recognition system. A global face detection and recognition system was also trained for comparison. Results on real test images show that the component-based recognition system clearly outperforms a comparable whole face recognition system. Component-based techniques yielded a recognition rate of 90% for faces rotated to approximately to ±45◦ in depth under varying illumination conditions. In comparison, global techniques only performed at a recognition rate of 40% on the same test set. These results point to the overall robustness of component-based face recognition in comparison with global recognition.

REFERENCES

[1] B. Heisele, T. Serre, M. Pontil, and T. Poggio. Component-based face detection. In Proc. IEEE Conference on Computer Vision and Pattern Recognition, volume 1, pages 657–662, Hawaii, 2001.

[2] B. Moghaddam, W. Wahid, and A. Pentland. Beyond eigenfaces: probabilistic matching for face recognition. In Proc. IEEE International Conference on Automatic Face and Gesture Recognition, pages 30–35, 1998.

[3] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. In IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 23, pages 349–361, April 2001.

[4] Ayache, N. and O. D. Faugeras: 1986, ‘Hyper: a new approach for the recognition and positioning of two-dimensional objects’. IEEE Transactions on Pattern Analysis and Machine Intelligence 8(1), 44–54.

[5] Baker, S. and T. Kanade: 2002, ‘Limits on Super-Resolution and How to Break Them’. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(9), 1167–1183.

[6] A. Aner and J. R. Kender. Video summaries through mosaic-based shot and scene clustering. In European Conference on Computer Vision, pages 388–402, Copenhagen, Denmark, 2002.

[7] N. Ayache and O. D. Faugeras. Hyper: a new approach for the recognition and positioning of two-dimensional objects. IEEE Transactions on Pattern Analysis and Machine Intelligence, 8(1):44–54, January 1986.

[8] A. Baumberg. Reliable feature matching across widely separated views. In Conference on Computer Vision and Pattern Recognition, pages 774–781, 2000.

[9] J. B. Burns, R. S. Weiss, and E. M. Riseman. View variation of point-set and line-segment features. IEEE Transactions on Pattern Analysis and Machine Intelligence, 15(1):51–68, January 1993.

[10] O. D. Faugeras, Q. T. Luong, and T. Papadopoulo. The Geometry of Multiple Images. MIT Press, 2001.