Modeling 3D Objects from Stereo Views and Recognizing Them in Photographs

Akash Kushal[1] and Jean Ponce[1,2]

[1] Department of Computer Science,
University of Illinois at Urbana Champaign, USA
[2] Département d'Informatique,
Ecole Normale Supérieure, Paris, France
{kushal, ponce}@cs.uiuc.edu

Abstract. Local appearance models in the neighborhood of salient image features, together with local and/or global geometric constraints, serve as the basis for several recent and effective approaches to 3D object recognition from photographs. However, these techniques typically either fail to explicitly account for the strong geometric constraints associated with multiple images of the same 3D object, or require a large set of training images with much overlap to construct relatively sparse object models. This paper proposes a simple new method for automatically constructing 3D object models consisting of *dense* assemblies of small surface patches and affine-invariant descriptions of the corresponding texture patterns from *a few* (7 to 12) stereo pairs. Similar constraints are used to effectively identify instances of these models in highly cluttered photographs taken from arbitrary and unknown viewpoints. Experiments with a dataset consisting of 80 test images of 9 objects, including comparisons with a number of baseline algorithms, demonstrate the promise of the proposed approach.

1 Introduction

This paper addresses the problem of recognizing three dimensional (3D) objects in photographs taken from arbitrary viewpoints. Recently, object recognition approaches based on local viewpoint invariant feature matching ([1], [2], [3], [4]) have become increasingly popular. The local nature of these features provides tolerance to occlusions and their viewpoint invariance provides tolerance to changes in object pose. Most methods (for example [5],[6]) match each of the training images of the object to the test image independently and use the highest matching score to detect the presence/absence of the object in the test image. This essentially reduces object recognition to a wide-baseline stereo matching problem. Only a few previous approaches ([2], [7], [8]) exploit the relationships among the model views. Lowe [2] clusters the training images into model views and links matching features in adjacent clusters. Each test image feature matched to some feature f in a model view v votes for v and its neighbors linked to f. This helps to model feature appearance variation since different model views provide slightly different pictures of the features they share, yet features' votes do not

A. Leonardis, H. Bischof, and A. Pinz (Eds.): ECCV 2006, Part II, LNCS 3952, pp. 563–574, 2006.

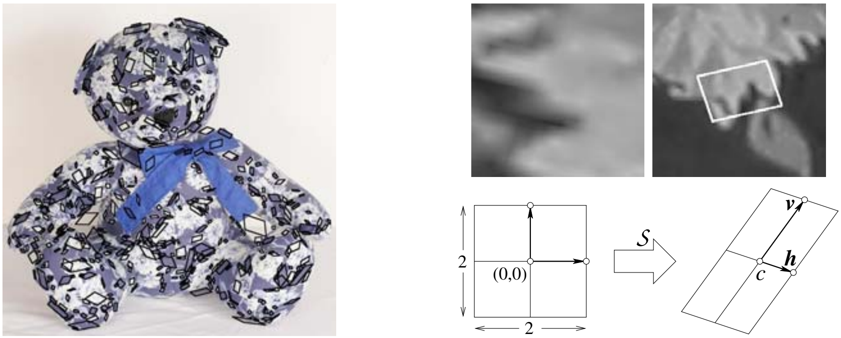
get dispersed among competing model views. Ferrari *et al.* [7] integrate the information contained in successive images by constructing *region tracks* consisting of the same region of the object seen in multiple views. They introduce the notion of a group of aggregated matches (GAM) which is a collection of matched regions on the same surface of the object. The region tracks are then used to transfer matched GAMs from one model view to another, and their consistency is checked using a heuristic test. The problem with this (as with all other methods that do not explicitly exploit 3D constraints) is that geometric consistency can only be *loosely* enforced. Also, for both [2] and [7] there is no way to determine consistency among matched regions which are not seen together in any model view. Rothganger *et al.* [8] use multiple images to build a model encoding the 3D structure of the object, and the much tighter constraints associated with the 3D projection of the model patches are used to guide matching during recognition. In this case, the 3D model explicitly integrates the various model views, but the determination of the 3D position and orientation of a patch on the object requires it to be visible in three or more training images [8], and hence requires a large number of closely separated training images for modeling the object. Also, [8] only makes use of patches centered at interest points, so the model constructed is sparse and does not encode all the available information in the training images. We tackle these issues by using calibrated stereo pairs to construct partial 3D object models and then register these models together to form a full model.[1] This allows the use of a sparse set of stereo training views (7 to 12 pairs in our experiments) for the modeling. We also extend to 3D object models the idea proposed in [6] in the image matching domain, and augment the model patches associated with interest points of [8] (called *primary* patches from now on) with more general *secondary* patches. This allows us to cover the object densely, utilize all the available texture information in the training images, and effectively handle clutter and occlusion in recognition tasks.

The paper is organized as follows. In section 2 we discuss the detection and representation of affine invariant patches as well as give an overview of our approach. The construction of the partial models and their inter-registration to generate the full model is explained in section 3. The details of the recognition phase of the algorithm are provided in section 4. In section 5 we show recognition results using the proposed approach and summarize in section 6.

2 An Overview of Our Approach

We use an implementation of the *affine region* detector developed by Mikolajczyk and Schmid [3] for low-level image description. The detector is initialized with the Harris-Laplacian interest point detector and the Difference of Gaussian (DoG) operator similar to [9]. The two detectors find complementary type of points. The Harris-Laplacian detector tends to find corners and points at which significant intensity changes occur while the DoG operator detects blob like features in the image. The output of the interest point detection/rectification process is

[1] This is for modeling only of course; individual photographs are used for recognition.



(a) Affine regions found in an image of a teddy bear. Only a subset of the patches detected is shown for clarity.

(b) The inverse transformation S maps the rectified square associated with an affine region back onto the image

Fig. 1. Affine regions and inverse rectification

a set of parallelogram-shaped image patches together with the corresponding affine *rectifying* transformations mapping these onto a square with edge length 2 centered at the origin. We represent each detected region by the 2×3 affine transformation matrix $\mathcal{S}$ that maps the rectified texture patch back onto its position in the image as shown in Fig. 1(b) (after [8]).

We use calibrated stereo for determining the 3D structure of the object and building the model mentioned in the previous section. Potential *primary matches* between the affine regions found in each stereo pair are first filtered using photometric and geometric consistency constraints, and then augmented with additional *secondary matches* for dense coverage of the object, as proposed in [6] in the 2D case. The 3D location and shape of the patches is determined using standard stereo to generate partial models which are later combined to form a complete model of the object. The 3D patches that correspond to primary (or secondary) matches are called primary (or secondary) model patches.

A similar scheme is followed during recognition. First, the primary patches in the model are matched to the affine regions found in the test image. These primary patches are then used as guides for matching nearby secondary patches. The object is recognized based on the number of matched patches.

3 Stereo Modeling

We start by acquiring a few (7 to 12) stereo pairs that are roughly equally spaced around the equatorial ring of the object for modeling. The stereo views are taken against a uniform background to allow for easy segmentation. Then, a standard stereo matching algorithm that searches for matching patches along corresponding epipolar lines is used to determine an initial set of tentative matches. We use a combination of SIFT [5] and the color histogram descriptor described in [10] to compute the initial matches. The matches are then refined to obtain the correct alignment of the patches in the left and right images. Only matches with normalized correlation greater than a pre-refinement threshold (kept at 0.75) are considered for the refinement step for efficiency reasons. The refinement process employs nonlinear optimization to affinely deform the right image patch until the correlation with its match in the left image is maximized. Matches with normalized correlation greater than a post-refinement threshold (equal to 0.9 in this paper) are kept for subsequent processing.

The matches are filtered by using a neighborhood constraint which removes a match if its neighbors are not consistent with it. More precisely, for every match m we look at its K closest neighbors in the left image ($K = 5$ in our implementation) and, for every triple out of these, we calculate the barycentric coordinates of the center of the left and right patches of m with respect to the triangle formed by the centers of the patches of the triple in the left and right images respectively. We then count the number of triples for which these barycentric coordinates agree (the sum of squared differences is smaller than a tolerance limit $\mathcal{L} = 0.5$). We repeat the process using the K closest neighbors of m in the right image and add up both the counts. Finally, the matches with a count smaller than a threshold T are dropped. Setting $T = 2\binom{K-1}{3}$ ensures that a correct match with one bad nearby match out of the K still survives after this test. This gives us a set Γ of reliable matches. Note that these matches are based only on the primary patches associated with salient affine regions detected in the stereo training images and hence, only cover the object sparsely. To get a dense coverage of the object we use an expansion technique similar to [6] to spread these initial matches in Γ.

Expansion Technique

We use the fact that the training views are taken against a uniform background to segment the object and cover it with a grid Ω of partially overlapping square-shaped patches in the left image (Fig. 2(a)). For every match m_i in Γ, we

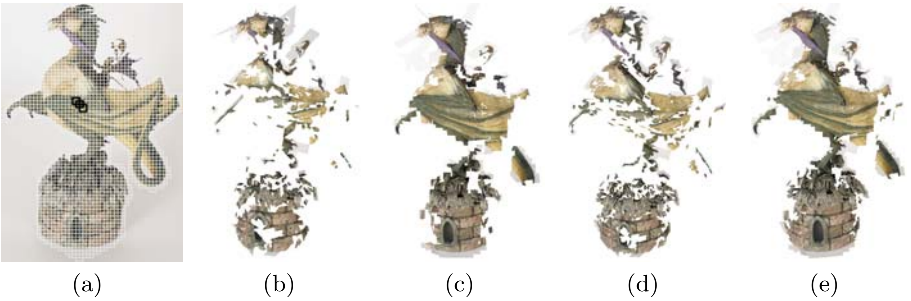

(a) (b) (c) (d) (e)

Fig. 2. (a) Left image in a stereo pair, covered with a grid of patches (three of the overlapping patches are shown in black for clarity). (b) Partial model constructed from primary matches before expansion. (c) Model constructed using only the secondary patches found during expansion. (d) Model containing the primary patches after expansion. (e) Model containing all the patches after expansion.

compute the affine transformation $\mathcal{T} = \mathcal{S}_{R_i}\mathcal{S}_{L_i}^{-1}$ between the corresponding patches L_i and R_i in the left and right images. Here $\mathcal{S}_{L_i}$ and $\mathcal{S}_{R_i}$ are the inverse rectification matrices for L_i and R_i respectively. We use $\mathcal{T}$ to predict the location $\mathcal{S}_{R_j} = \mathcal{T}\mathcal{S}_{L_j}$ of the right matches of the yet unmatched patches L_j in Ω that are close to (within one side length of) the center of L_i. Then, a refinement process is used to align the predicted patch correctly in the right image. Again, if the match has sufficient correlation after refinement, it is accepted as a valid match and added to Γ. Since the patches that form these matches are not associated with interest points, we call these secondary matches. The expansion process iterates by expanding around the newly added matches to Γ until no more matches can be added. This process usually covers the entire object surface densely with matches. Figure 2(c) shows the secondary patches on a partial model of the dragon constructed from a single stereo pair.

We then use the secondary matches to locate additional primary matches associated with salient affine regions. Even though the corresponding part of the object surface may already be covered (with secondary matches), this is useful because it is the primary matches that can be repeatably detected, and will later be required for the initial matching to the test image as well as for the alignment of the partial models. This is accomplished by finding unmatched affine regions in the left (respectively right) image, and using close-by secondary matches to predict the position of the corresponding patches in the right (respectively left) image. Again, a refinement process is used to adjust the alignment of the right (respectively left) image patch. If there is sufficient correlation (again 0.9) between the left and right patches, the match is added to Γ. Figures 2(d) and 2(e) respectively show the expanded primary patches and the union of the primary and secondary patches in the partial model of the dragon.

Model Construction

The dense matches constructed as discussed above are used for building the 3D model. First, we solve for the patch centers in 3D by using standard calibrated stereo triangulation. Then, we reconstruct the edges of the corresponding parallelograms using a first-order approximation to the perspective projection equations in the vicinity of the patch centers as proposed by Rothganger [10]. This gives us a partial 3D model of the object for each stereo pair. The next task is to combine these partial models into a complete model.

The first step in combining the models is to find appearance-based matches between the primary model patches in adjacent partial models. Again, SIFT and color histogram descriptors are used to facilitate the initial matching. Next, a variant of the expansion scheme described earlier is used to propagate these initial matches between 3D patches to neighboring model patches as follows (Fig. 3). Let the two partial models being registered be M_P and M_Q. For each initial match $\mathcal{M}_i$ between the 3D patches $\mathcal{P}_i$ in M_P and $\mathcal{Q}_i$ in M_Q, we consider the 2D patch p_i (resp. q_i) corresponding to $\mathcal{P}_i$ (resp. $\mathcal{Q}_i$) in the left stereo image of M_P (resp. M_Q). We calculate the affine transformation $\mathcal{T}$ that maps the patch p_i onto q_i. Then, we consider the yet unmatched patches P_j in M_P whose 2D

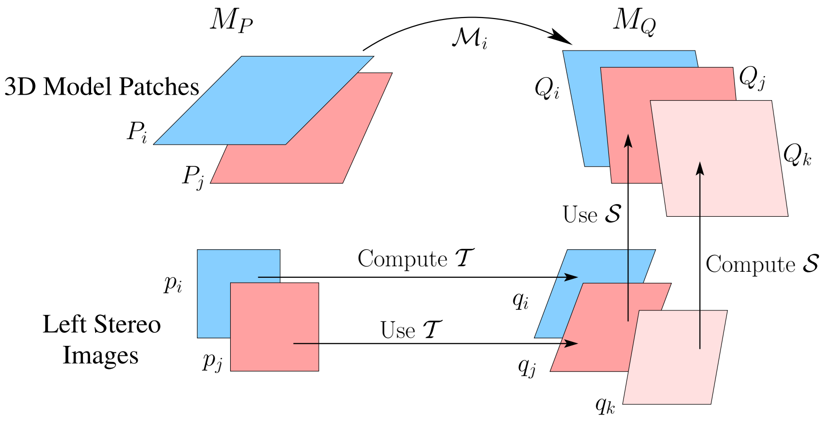
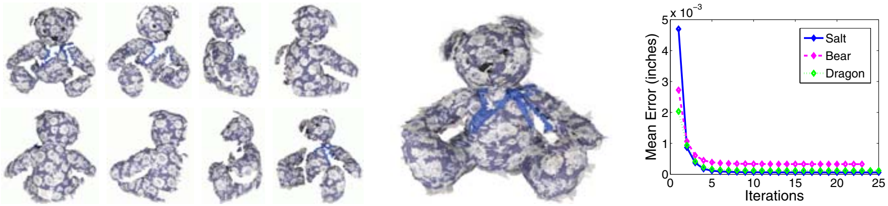


Fig. 3. Expansion during registration

projection p_j in the left stereo image lies within a small distance limit of the center of p_i. These patches p_j are then projected to q_j in the left stereo image of M_Q using $\mathcal{T}$. A refinement process (similar to the one described earlier) is then used to align the projected patch q_j correctly. The match is removed from consideration if the final correlation between p_j and q_j's normalized representation is less than a threshold (again kept at 0.9). If the match passes this test we find the patch Q_k in M_Q whose projection q_k into the left stereo image of M_Q is closest to q_j's center point. An estimate of the position of the 3D patch Q_j that corresponds to the 2D patch q_j can then be obtained, assuming that Q_j lies on the same plane π_k as Q_k. An affine transformation $\mathcal{S}$ that maps the 2D patch q_k to the 3D patch Q_k on π_k is calculated and then Q_j is estimated by projecting q_j onto π_k using $\mathcal{S}$. This new match between P_j and Q_j is then added to the set of matches and is used for finding other matches. This expansion step has proven to be very useful while registering models with small overlap.

Finally, all the matches are filtered through a RANSAC procedure that finds the matches consistent with a rigid transformation. This provides an estimate of the pairwise rigid transformations. Since these pairwise estimates may not in general be consistent with each other (the product of the rotations between the consecutive models must be the identity), we use a process similar to [11] to find a consistent solution: It is initialized using the pairwise transformation estimates and these estimates are refined by looping through all the partial models and updating the position of the current model to align it best with its neighbors. More formally, we search for the rigid transformation that minimizes the sum of squared distances between the centers of the matched patches in the current model and its neighbors. The positions of these neighbors are kept fixed while the position of the current partial model is calculated via linear least squares [11]. The above process is iterated until a local minimum of the error is reached. Figure 4(c) shows a plot of the mean squared error after each iteration of the refinement process for three of the models used for experimentation. Finally the rigid transformations estimated are used to bring all the partial models into a



(a) Partial models (b) Complete model (c) Mean error

Fig. 4. Registration of partial models

common euclidean coordinate frame and a complete model is constructed by taking the union of these transformed partial models. The partial models and the complete model formed after registration for a teddy bear are shown in Figs. 4(a) and 4(b) respectively.

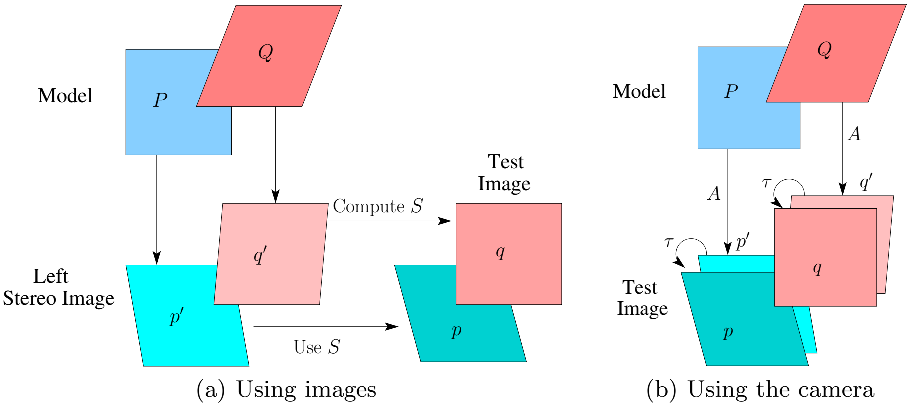
4 Recognition

The recognition starts by matching the repeatable primary patches in the 3D model to the interest points detected in the test image. Again, we use both SIFT descriptors and color histograms to characterize the appearance of the patches and compute the initial matches. The refinement process is then employed to affinely deform the matched test image patch so as to maximize the its correlation with its corresponding model patch. Matches with correlation smaller than a threshold (again taken as 0.9) are dropped before further processing. The remaining matches are used as seeds for the subsequent match expansion stage.

Expansion Process

This process is similar in spirit to the expansion technique used during the initial modeling but the expansion here happens on the surface of the 3D model instead of the stereo images. For this, we first preprocess the model M to build an undirected graph G_M that represents the adjacency information of the patches in M. We add an edge e between two patches if their centers lie within a distance limit of each other. This limit is set to be such that the average degree of a vertex is around 20. We now spread the matches along the edges of this graph using the following steps.

Expansion using images (Fig. 5(a)): This expansion step is similar to the expansion during modeling. For each previously matched model patch P we calculate the affine transformation $\mathcal{S}$ that maps its projection in the left training image of the stereo pair from which it originates into the test image. Then we look at every unmatched neighbor Q of P that is part of the same partial model (and so shares the same left stereo image) and use $\mathcal{S}$ to predict its location in the test image. This predicted position is then refined as before and the match is accepted if the correlation is sufficiently large (again 0.9). This expansion scheme does not allow expanding matches from one partial model to another.



(a) Using images (b) Using the camera

Fig. 5. Expansion during recognition

Geometric consistency test (Algorithm 1): A *"greedy"* RANSAC-like algorithm is used to extract a set of geometrically consistent matches and estimate the camera for the test image. The test image camera is modeled as a weak-perspective camera with zero skew and square pixels.

Expansion using the camera (Fig. 5(b)): This step is used after the matches have been filtered through the above geometric consistency test and the camera A associated with the test image has been estimated. A is used to project a base 3D patch P (which is already matched to a patch p in the test image) and some adjacent patch Q into the test image. Let the 2D projected patches be p' and q' respectively. A correcting affine transformation τ is computed that aligns the projection p' of the base 3D match exactly with its correct location p. τ is then applied to the projection q' of the adjacent patch to obtain a corrected prediction q of its position. The prediction is then refined as before to maximize the normalized correlation between the patches corresponding to the match and accepted only if it has high correlation (greater than 0.9). This expansion step allows for moving smoothly from one partial model to another and hence provides an advantage over the pure 2D expansion technique of [6].

For extending matches to parts of the object that are not directly connected to the initial matches in the test image (possibly due to occlusion) the reconstructed test camera is used to project unmatched primary patches from the model into the test image. Affine regions detected in the test image close to these projected positions are then matched to the corresponding model patch. Again, the refinement process is used to correctly align the patch in the test image and the match is accepted if the correlation exceeds a threshold (again, 0.9).

The two expansion steps also allow us to reject false matches by simply removing those that do not have enough support. More precisely, if the expansion step from a base match tries to expand to a large number of neighbors and none of

```
Input: A set $M$ of possible matches.
Output: A set $S$ of trusted matches, camera for the test image $C$
  for $i = 1$ to $numIter$ do
    • Pick a match $m_i \in M$ at random.
    • Select the most compatible match $m'_i \in M \setminus \{m_i\}$ to $m_i$.
    • Initialize $S_i = \{m_i, m'_i\}$ and $C_i$ to the camera estimated using $S_i$.
    • Select $m_{best} \in M \setminus S_i$ with minimum reprojection error $\mathcal{E}_{best}$ using $C_i$
    while $|S_i| < K$ and $\mathcal{E}_{best} < \tau$ do
      • $S_i \Leftarrow S_i \cup \{m_{best}\}$.
      • Update $C_i$ with the camera estimated using $S_i$
      • Select $m_{best} \in M \setminus S_i$ with minimum reprojection error $\mathcal{E}_{best}$ using $C_i$
    end while
    • Add all matches $m \in M \setminus S_i$ with reprojection error $\mathcal{E}_m < \tau$ to $S_i$.
  end for
  • Set $S$ to the $S_i$ with the largest cardinality.
  • Estimate the camera $C$ for the test image using $S$.
```

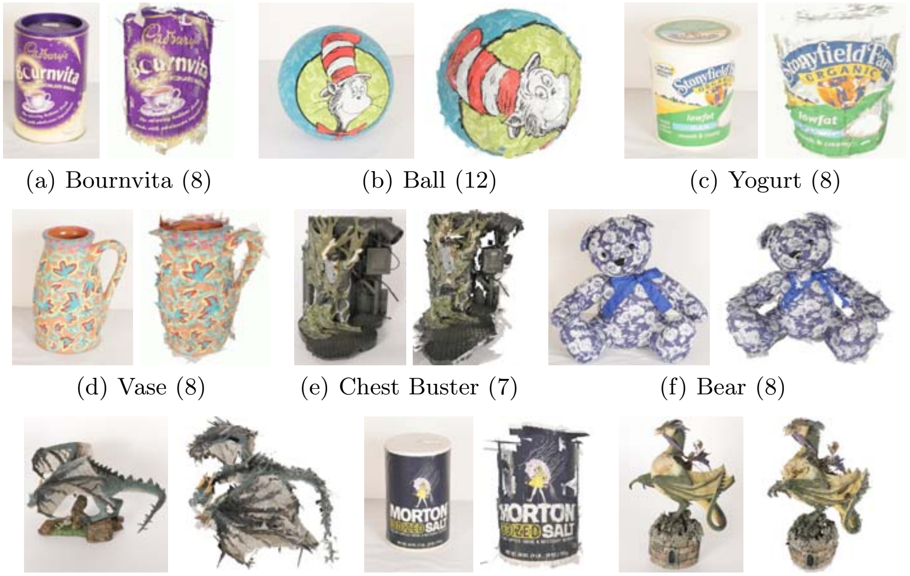
Algorithm 1. Geometric consistency test

these succeeds in forming an acceptable match, the base match is removed. The above cycle consisting of the two expansion steps and the geometric consistency test is iterated until the number of matches does not increase any more. This process usually takes only 3 iterations.

5 Results

We have evaluated the proposed method on a dataset consisting of 9 objects and 80 test images. The object models, constructed from 7 to 12 stereo views each, are shown in Fig. 6. The objects vary from simple shapes (e.g., the salt container) to quite complex ones (e.g., the two dragons and the chest buster model).

The test images contain the objects in different orientations and under varying amounts of occlusion and clutter. The total number of occurrences of the objects in the test image dataset is 129 since some images contain more than one object. Figure 7(a) shows the ROC plot between the true positive (detection) rate and the false positive rate. To assess the value of the expansion step of our approach, we have simply removed the secondary patches and the extra primary patches added during this stage of modeling from our models, and used these sparse models for recognition (this is similar in spirit to the algorithm proposed by Rothganger et al. [8], but includes the expansion step during the recognition phase which was absent in [8]). The corresponding recognition performance is depicted by the blue ROC curve. Our experiments clearly demonstrate the benefit of using dense models as opposed to sparse ones for our dataset. We have also implemented recognition as wide-baseline stereo matching to assess the power of using explicit 3D constraints as opposed to simple epipolar ones. Each test image is matched to all the 168 training images (both left and right images for each stereo pair) for every object separately, making a total of $168 \times 80 = 13440$

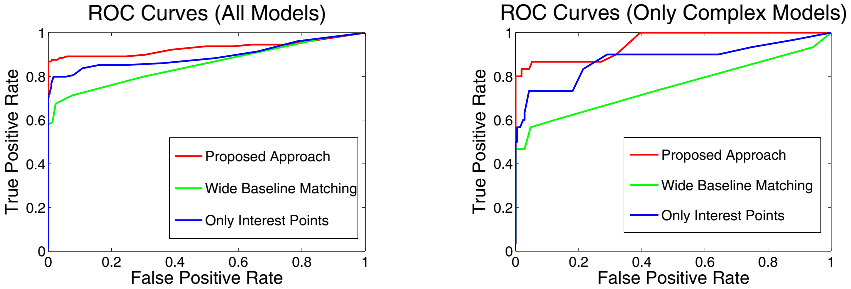


(a) Bournvita (8) (b) Ball (12) (c) Yogurt (8)

(d) Vase (8) (e) Chest Buster (7) (f) Bear (8)

(g) Small Dragon (12) (h) Salt (8) (i) Dragon (12)

Fig. 6. Object models. The number of stereo views used is given in parenthesis.



(a) ROC (all models) (b) ROC (Chest Buster & both Dragons)

Fig. 7. Comparison ROC plots

image pairs to be compared. The maximum number of matches corresponding to each object is recorded and used to construct the ROC curve. As expected, our method clearly outperforms this simple baseline approach. The detection rates for zero false positives and the equal error rates for the different methods are shown in Fig. 5.

The proposed approach also performs well on the highly complex geometric objects like the dragons and the chest buster model. Figure 7(b) shows the comparison of the ROC plots on the dataset restricted to only these 3 models. The variation in appearance of the features due to small viewpoint changes is

Method	Detection Rate (zero false positives)	Equal Error Rate
Proposed Approach	86.8%	89.1%
Primary patches only	69.8%	84.9%
Wide Baseline	58.1%	77.1%

Fig. 8. Error rate comparison

larger for these models since the surface of the models is not smooth. Because the proposed approach combines the different views of the features together (when the different partial models are merged) its performance is less severely affected on the restricted dataset. On the other hand, the performance of the wide-baseline matching scheme drops by a significant amount.

Finally, Fig. 9 gives a qualitative illustration of the performance of our algorithm with a gallery of recognition results on some test images which contain the objects under heavy occlusion, viewpoint and scale variation, as well as extensive clutter. The dataset used in this paper is available at the following URL: `http://www-cvr.ai.uiuc.edu/ponce_grp/data/stereo_recog_dataset/`

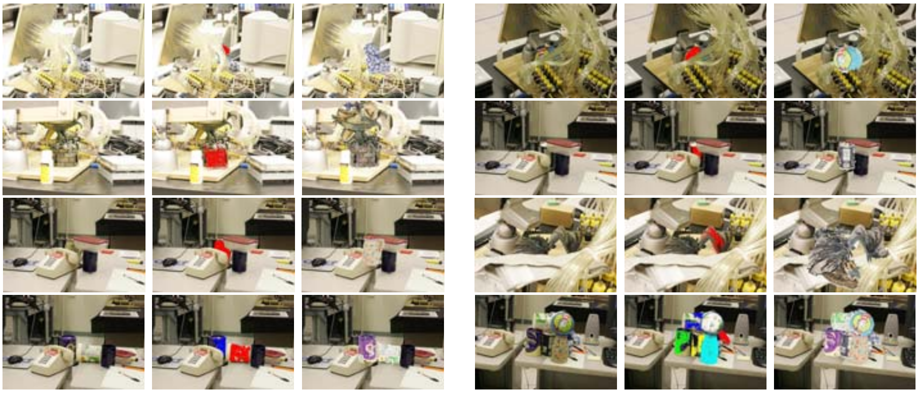


Fig. 9. Results: test image (left), matched patches (center), predicted location (right)

6 Conclusions and Summary

We have proposed an approach to efficiently build dense 3D euclidean models of objects from stereo views and use them for recognizing these objects in cluttered photographs taken from arbitrary viewpoints. At this point there are many directions for future work.

- Extending the approach to handle non rigid deformations
- Recognizing objects in a cluttered scene using a pair of calibrated stereo images of the scene.

Also, it would be desirable to do a comparison with the native implementations of other state-of-the-art recognition methods such as those proposed by Ferrari et al. [7], Lowe [2], and Rothganger et al. [8].

Acknowledgments

This research was supported in part by the National Science Foundation under grants IIS 03-12438, IIS 03-08087 and IIS 05-35152 and the Beckman Institute.

References

1. Tuytelaars, T., Van Gool, L.J.: Content-based image retrieval based on local affinely invariant regions. In: Visual Information and Information Systems. (1999) 493–500
2. Lowe, D.G.: Local feature view clustering for 3d object recognition. In: Conference on Computer Vision and Pattern Recognition. (2001)
3. Mikolajczyk, K., Schmid, C.: An affine invariant interest point detector. In: European Conference on Computer Vision. Volume I. (2002) 128–142
4. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: British Machine Vision Conference. Volume I. (2002) 384–393
5. Lowe, D.G.: Object recognition from local scale-invariant features. In: International Conference on Computer Vision, Corfu, Greece (1999) 1150–1157
6. Ferrari, V., Tuytelaars, T., Van Gool, L.: Simultaneous object recognition and segmentation by image exploration. In: European Conference on Computer Vision. (2004)
7. Ferrari, V., Tuytelaars, T., Gool, L.V.: Integrating multiple model views for object recognition. In: Conference on Computer Vision and Pattern Recognition. (2004)
8. Rothganger, F., Lazebnik, S., Schmid, C., Ponce, J.: 3d object modeling and recognition using affine-invariant patches and multi-view spatial constraints. In: Conference on Computer Vision and Pattern Recognition. Volume II. (2003) 272–277
9. Rothganger, F., Lazebnik, S., Schmid, C., Ponce, J.: 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. International Journal of Computer Vision (In press, 2005)
10. Rothganger, F.: 3D object modeling and recognition in photographs and video. PhD thesis, University of Illinois, Urbana Champaign (2004)
11. Raouf Benjemaa, F.S.: A solution for the registration of multiple 3d point sets using unit quaternions. In: European Conference on Computer Vision. (1998)