3D Object detection and viewpoint selection in sketch images using local patch-based Zernike moments

Anh-Phuong Ta Christian Wolf Guillaume Lavoué Atilla Baskurt
Université de Lyon, CNRS
INSA-Lyon, LIRIS, UMR5205, F-69621, France
{anh-phuong.ta,christian.wolf,guillaume.lavoue,atilla.baskurt}@liris.cnrs.fr

Abstract

In this paper we present a new approach to detect and recognize 3D models in 2D storyboards which have been drawn during the production process of animated cartoons. Our method is robust to occlusion, scale and rotation. The lack of texture and color makes it difficult to extract local features of the target object from the sketched storyboard. Therefore the existing approaches using local descriptors like interest points can fail in such images. We propose a new framework which combines patch-based Zernike descriptors with a method enforcing spatial constraints for exactly detecting 3D models represented as a set of 2D views in the storyboards. Experimental results show that the proposed method can deal with partial object occlusion and is suitable for poorly textured objects.

Keywords: 2D/3D object detection, localization and recognition, pose recognition, rotation angle retrieval, local features

1. Introduction

In the process of creating 3D animated films, the manual creation of 3D scenes out of storyboards sketched by artists is a tedious process, hence the desire to perform it automatically with a computer vision algorithm[1]. Indeed, currently, the storyboards are hand drawn by the artists in charge of the scenario of the movie, mostly with traditional non electronic pens. Then, for each episode, modeling specialists create the 3D representation of the different scenes based on the storyboards by using existing 3D object models stored in a database.

In this paper we address the goal of recognizing each 3D model from the corresponding piece of sketch, along with its 3D viewing angle, its scale and its rotation angle in the drawing plane, so that it can be automatically and correctly placed in the 3D scene.

Object recognition is a fundamental problem in computer vision. Recognizing three-dimensional objects in a 2D scene is a well known problem. On direct 3D object recognition, very few work exists, e.g [7]. We chose to tackle the problem by recognizing a 3D object through a set of 2D images each corresponding to a single viewpoint. In the state of the art, we can find a great variety of 2D object recognition methods, some successful methods are those of Lowe [11], Belongie et al. [2], Furgus et al. [5], Shotton et al. [14], Opelt et al. [12]. However, most of these methods work only on textured objects and fail on smooth shapes, such as sketch images (drawings). In such cases, global Zernike moments are particularly robust; they have been successfully used for 2D/3D object recognition through sketches in [1][8][13].

One of the main difficulties is the frequent occlusion problem, which in image indexing and in object detection is commonly tackled using local features [9][4]. In these local approaches, local descriptors are calculated on keypoints [11][17] or on edges[6]. These methods are therefore not applicable for sketch retrieval.

However, interest points being very unstable on sketch images[2], these methods are not applicable in our case.

Inspired from Revaud et al's work [13], we chose Zernike moment descriptors. However, we employ them in a local manner, which allows us to overcome the occlusion problems. In contrast to early local approaches presented above, we introduce a new patch-based Zernike moment method, which takes into account not only the features themselves, but also their local relationships. Three constraints are taken into account for the detection:

I. The Zernike moment distance between the model and the storyboard patch.

[1]This work has partly been financed by Pinka Productions (`http://www.pinka-prod.com`)

[2]This is confirmed by our experiments, in which we have attempted to use SIFT keypoints and descriptors for sketch recognition with very poor results.

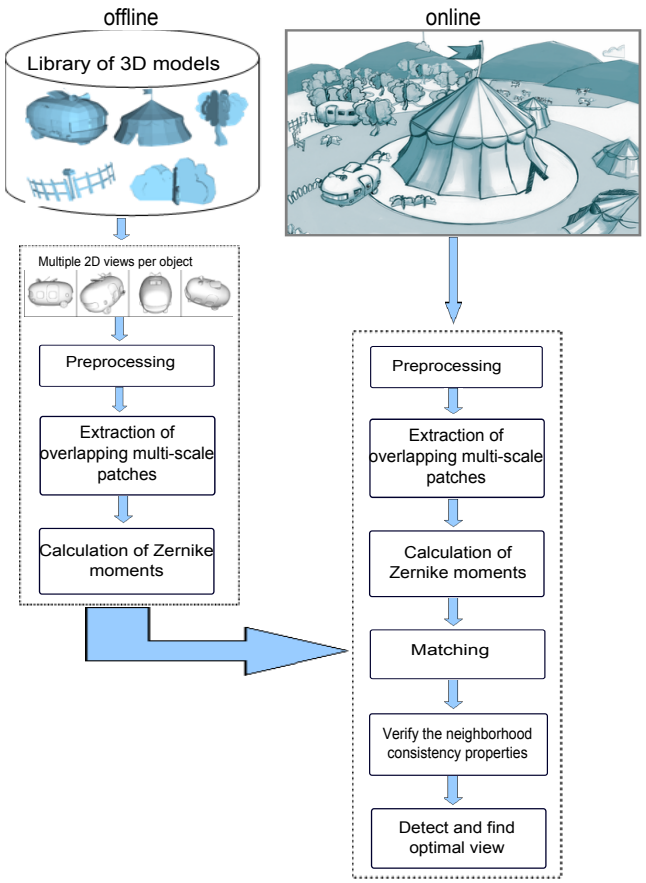


Figure 1: The proposed 3D patch-based object detection in sketch images.

II. The consistency of the neighborhood relationships between model and storyboard patches.

III. The consistency of the rotation angle among the neighboring patches.

Note that our objectives are two-fold:

- Detecting and recognizing the 3D models as well as their location and size
- Recognizing the 3D pose: detecting the viewpoint for each model

In our work, the latter viewpoint will be obtained by selecting the correct 2D view stored in the database as well as retrieving the in-plane rotation angle of this view.

The paper is organized as follows: the Zernike moment descriptor is reviewed in section 2. In section 3, we present our method. In section 4, we present experimental results illustrating the effectiveness of our method. Finally, we conclude and give some perspectives in section 5.

2. Zernike moment descriptors

Zernike moments have been widely used as features for pattern recognition. They are constructed by using a set of complex polynomials which form a complete orthogonal basis set defined on the unit disc. The Zernike moments are defined with an order p and a repetition q over $D = \{(p,q)|0 \leq p \leq \infty, |q| \leq |p|, |p-q| = even\}$,

$$Z_{pq} = \frac{p+1}{\pi} \int\int_{x^2+y^2 \leq 1} {V^*}_{pq} f(x,y) \partial x \partial y \qquad (1)$$

where $V_{pq}(x,y) = V_{pq}(\rho\cos\theta, \rho\sin\theta) = Z_{pq}(\rho)\exp(jm\theta)$, and

$$R_{pq}(\rho) = \sum_{k=|q|,|p-k|even}^{p} \frac{(-1)^{\frac{p-k}{2}} \frac{(p+k)}{2}!}{\frac{p-k}{2}! \frac{k-p}{2}! \frac{k+q}{2}!} \rho^k \qquad (2)$$

and ρ and θ are, respectively, the radius and the angle of the pixel (x,y) with respect to the object's gravity center.

To compute the distance between two Zernike descriptors, we use the comparator proposed by Revaud et al. [13], which returns both the distance (in the sense of similarity) and the rotation angle between the two patterns. This angle is further used for evaluating the rotation angle interaction of the neighboring patches in our method. In the following we will denote the Zernike distance between two patches as $d_{zd}(\cdot,\cdot)$ and the retrieved rotation angle as $d_{za}(\cdot,\cdot)$.

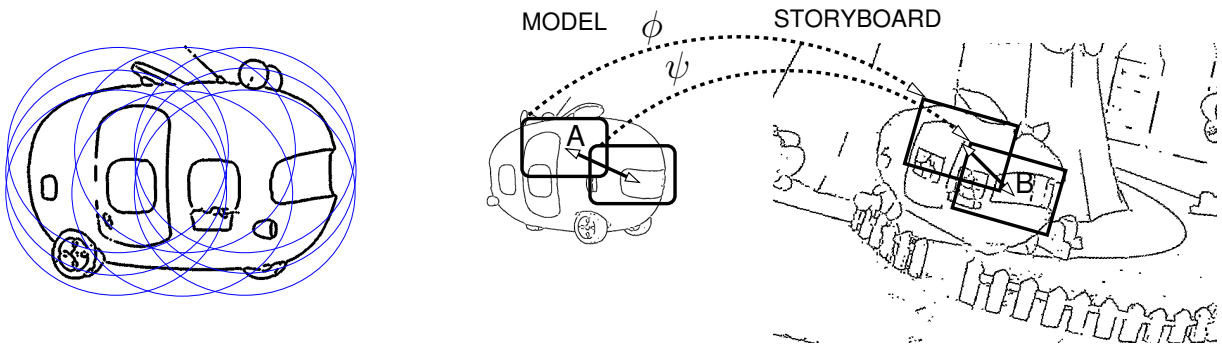
3. Proposed method

In figure 1, we present our general scheme for recognizing 3D objects in storyboards. Our method is based on the following principle: each 3D model is represented by a set of 2D views (model images); then, edges are detected in the model images with a canny detector – the storyboard images are already stroke images which do not necessite edge detection. All images are thresholded before subsequent preprocessing.

Patches of different sizes are then extracted from the model images as well as from the storyboard images (c.f. figure 2a). Each model patch is assigned to a storyboard patch by integrating the constraints described in section 1 (c.f. figure 2b).

In general, and for the full correspondence problem, there are M^S possible combinations of assignments, where M is the number of model patches and S the number of storyboard patches. For each of these assignments, $\approx M \cdot \bar{D}$ consistency criteria need to be checked, where $\bar{D}$ is the average number of neighbors of a patch. In general, this classification problem is NP-Hard [10].

Solutions for this problem including terms I. and II. above (without the rotation angle consistency; c.f section



(a) (b)

Figure 2: (a) Illustration of the overlapping patches extracted on a model view image: sizes range from 65% to 80% of the object size. (b) The different constraints: I. the Zernike distance assigns model patches to storyboard patches (dotted arrows); II. the euclidean distance A between neighboring patches is checked to be consistent with the euclidean distance B of neighboring storyboard patches; III. the rotation angle ϕ of one assignment is checked to be consistent with the rotation angle ψ of a neighboring assignment.

1) have been proposed: e.g by adopting the neighborhood structure in order to find the maximum *a posteriori* solution with the junction tree algorithm [15], which is of complexity $O(M \cdot S^4)$ and applicable to few primitives only; or by formulating the problem as a sampling procedure and checking for the consistency of each sample (RANSAC) [18, 3].

Instead of solving the full problem, we propose to proceed in two steps:

1. We assign model patches to storyboards patches using constraint I. only, i.e. the minimum Zernike distance (c.f. figure 2).

2. In a second step, we select among all model patches, the one which best verifies constraints II. and III.; we calculate a match score which is used for detection and viewpoint selection (c.f. figure 2).

The algorithm can be formalized as follows: we start from a list of model views $H = \{H_v\}(v = 1...V)$ and a list of N_v patches per view v: H_v=$\{h_{v1}, h_{v2}, ..., h_{vN_v}\}$, as well as a list of S scene patches C=$\{c_1, c_2, ..., c_S\}$.

As described above, in the first step we assign model patches to scene patches according to the minimal Zernike distance:

$$\forall v, \forall i : s_{vi} = \arg \min_{t \in [1..S]} d_{zd}(h_{vi}, c_t) \tag{3}$$

where s_{vi} denotes the storyboard patch assigned to model patch i of view v. Since the spatial relationships II. and III. are not used, this sub problem is of low complexity compared to the full problem.

In a second step, a match score $\alpha_{(v,i)}$ is calculated for each model patch h_{vi} for each view v; for each patch, we consider all its neighboring patches h_{vj}, i.e patches whose spatial euclidean distance is below a proximily threshold T_e:

$$\begin{aligned} \alpha_{(v,i)} = & \sum_{\substack{j:j\neq i, \\ d_e(h_{vi},h_{vj}) \leq T_e}}^{h_{vN_v}} \exp\left\{ -\frac{[d_e(h_{vi},h_{vj}) - d_e(c_{s_{vi}},c_{s_{vj}})]^2}{2\sigma_{de}^2} \right\} \\ & \times \exp\left\{ -\frac{[d_{za}(h_{vi},c_{s_{vi}}) \circ d_{za}(h_{vj},c_{s_{vj}})]^2}{2\sigma_{za}^2} \right\} \end{aligned} \tag{4}$$

where $d_e(h_{vi}, h_{vj})$ is the euclidean distance between the spatial positions of the model view patches h_{vi} and h_{vj}, d_{za} is the rotation angle between a model patch h_{vi} and its corresponding scene patch $c_{s_{vi}}$ and σ_{zd}, σ_{za} are variance parameters for Gaussian kernels parametrizing, respectively, the differences in euclidean distance and in Zernike angle. The first exponential measures the difference in spatial distance between two model patches and their corresponding scene patches. This distance should be close to zero if the object is subject to an isometry. The second exponential measures the consistency in the in-plane rotation angles between neighboring patches. The operator $\circ$ computes the difference between two angles taking into account their circular domain.

Finally, we calculate a single match value α_v^* for each view v by selecting the patch which best satisfies these geometrical constraints:

$$\alpha_v^* = \max_i \alpha_{(v,i)} \tag{5}$$

The view $\bar{\alpha}$ giving the highest score α is selected as the best view for the 3D model:

$$\begin{aligned} \alpha = & \quad \max_v \alpha_v^* \\ \bar{\alpha} = & \arg \quad \max_v \alpha_v^* \end{aligned} \tag{6}$$

The score value α is thresholded with a threshold T_d for detection.

Note that our algorithm not only detects the best view but also the in-plane rotation angle of the deteted viewpoint, which is based on the best detected patch.

In order to detect multiple objects, after a successful detection step we remove the corresponding patches from the storyboard and restart a new detection process.

4. Experiments

We applied our method to a real industrial application comprising five different 3D models: tents (2 different models), trailers, bushes, and trees provided by a company producing animated films. As shown in figure 1, the proposed method consists of two processing stages. For the offline process, about 120 views from each 3D model (i.e. $5 \times 120 = 600$ views on total) are extracted and then the views are indexed by calculating Zernike moments on 16 overlapping patches: the model patches are extracted on different spatial locations are of sizes varying from 65% to 80% of the model size. This boosts the probability of finding some of these patches even in the case of partial occlusion. In the online process, similar overlapping patches are extracted from the storyboard and Zernike moments are calculated at multiple-scales. We used 4 storyboards of size 3350×2260 for testing. We chose the same variance parameter which has been used in [15] for Gaussian kernels measuring differences in euclidean distances: σ_{de}=σ_{za}=0.4. The choice of an optimal parameter σ is far from trivial [16] and beyond the scope of this paper.

Figure 3 illustrates some results of our algorithm. Each object detected is marked with a bounding box and its corresponding 3D model view is displayed. Note the excellent results on the very difficult storyboard images. Most of the objects are occluded, some of them significantly, which does not prevent our method from correctly detecting and recognizing them. Furthermore, even very similar 3D models, as for instance the two different tents, are distinguished correctly.

We used the classical recall and precision measures for the evaluation of our method. Note that recall is defined as the amount of correctly detected objects with respect to the total amount of objects in the ground truth, whereas precision is the amount of correctly detected objects with respect to the total amount of detected objects. Since the two measures are dependent, i.e. varying the threshold T_d in order to increase recall will generally decrease precision, we chose to report the recall figures obtained for 100% precision, i.e. no false alarms.

Table 1 shows the comparison between Revaud et al's global approach [13] and our approach. The comparison results show that the proposed method attains higher recall

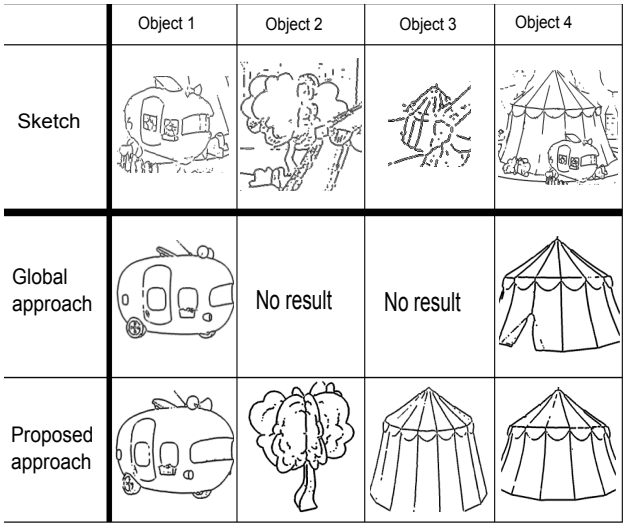


Figure 4: Examples of detection results, where the detected views are given for both compared methods.

	Global approach [13]		Proposed approach	
	total	%	total	%
tents	4/8	50.00	7/8	**88.00**
trailers	2/3	**67.00**	2/3	**67.00**
bushes	3/10	30.00	6/10	**60.00**
trees	1/31	3.00	4/31	**13.00**

Table 1: Comparison of recall for 100% precision for the global approach and the proposed approach.

than that of the global one. Both methods obtain low recall for the tree models. This is mainly caused by the fact that sometimes their sketch deviates too much from their 3D model. In addition, the sketched trees belong to highly cluttered scenes (forest $-$ see the tree models in figure 3). We do not consider the errors of the detected viewpoint in calculating recall and precision. A comparison of these errors for both methods is given in table 2.

Table 2 presents the mean error in viewpoint detection compared with the global method. We here show only the mean errors for 100% precision, which are calculated according to the results in table 1. Note that each view is characterized by two angles ($(\alpha \in [0, 2\pi], \beta \in [0, \pi])$. To compute the error of the viewpoint detection, we translate the angle pair into a corresponding 3D point on the unit sphere. The error between the detected viewpoint and the sketched viewpoint is then calculated as the euclidean distance between the corresponding 3D points on the sphere. The comparison results in table 2 demonstrate that our method performs better than the global one in viewpoint detection.

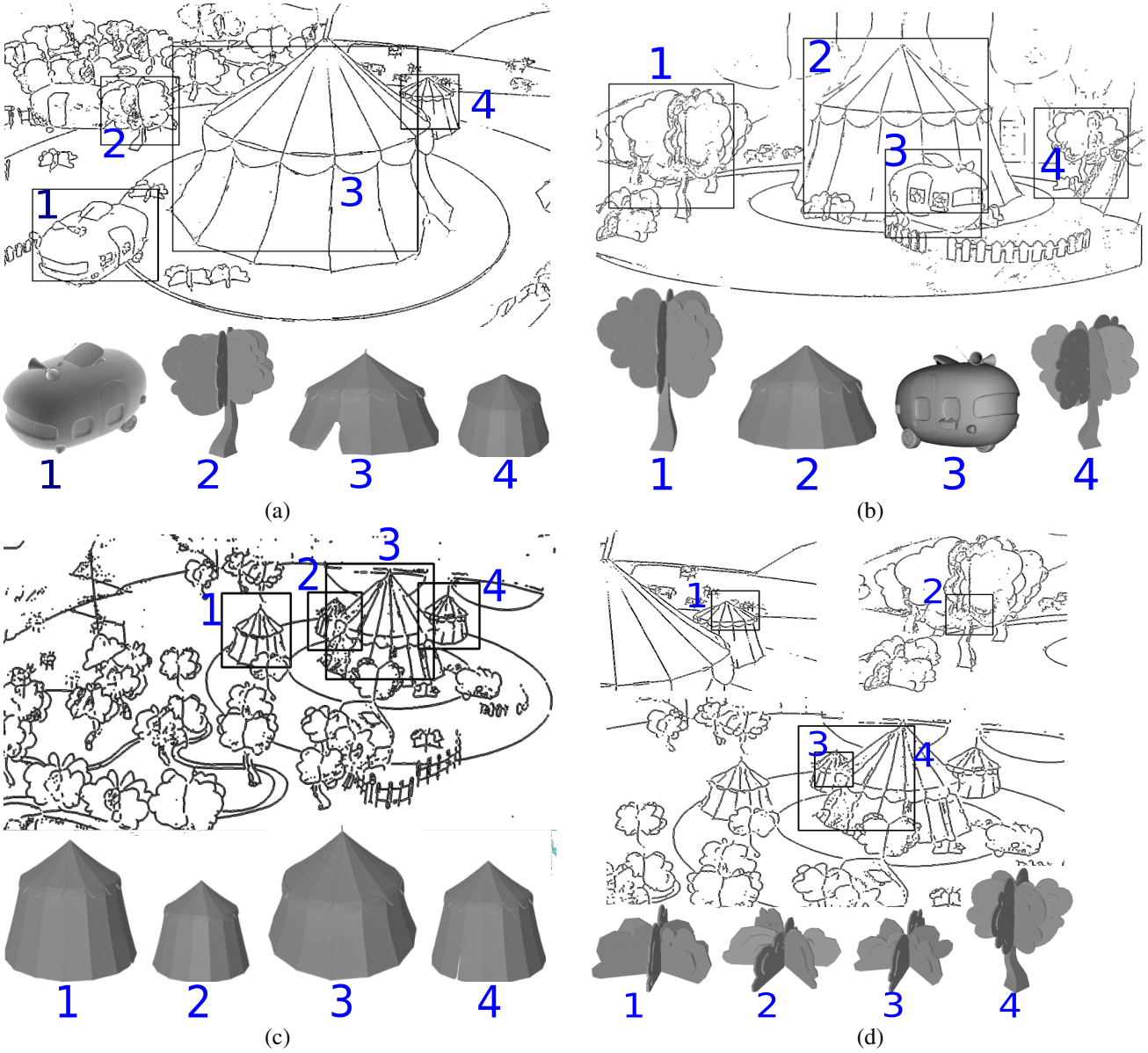


Figure 3: Examples of detection results on several storyboards. Note the successful detection in spite of many occlusions. Images (a)-(c) show detection results for 100% precision, i.e. no false alarms. Figure (d) illustrates the difficulty of the tree and bush models on an image created as a mixture of images (a)-(c). Searching for four tree and bush models, the best response for each detection are wrong models or parts of wrong models (tents etc.).

In figure 4 some visual results are presented for both approaches: in the first row are some sketched objects extracted from the storyboards, the second and third row show the detected views of the global approach and the proposed approach, respectively. We can see that the global approach is very sensitive to occlusions: for objects 1 and 4, which are slightly occluded, the global approach returned the correct object model but views which are not very similar to the views of the sketched objects, while our method can find the best views (the view which is closest to the sketched object in the database). For objects 2 and 3, which are more severely occluded, the global method failed, whereas our method can still recognize the 3D models although the detected views are sometimes slightly different from the original ones.

	Global approach [13]	Proposed approach
tents	0.81	**0.31**
trailers	0.29	**0.06**
bushes	**1.16**	1.41
trees	2.70	**1.36**

Table 2: Comparison of the mean error in viewpoint detection between the global approach and the proposed approach.

5. Conclusion

We have presented a new approach to recognize 3D models in storyboards. Our contribution is two-fold: First, we proposed a new local patch-based representation using Zernike moments, which is suitable for sketch recognition and is robust to occlusion, rotation and scale. Secondly, in addition to the Zernike distance, we have incorporated two main relationships between patches: spatial interactions and rotation angle consistency. This makes the proposed method flexible and easy to use with other descriptors.

Although only rigid transformations preserve distances, which are checked by our method, a very large class of non-rigid transformations is handled by the gausians kernels in equation (4) and the fact that interactions between neighbors are verified as opposed to more arbitrary interactions, as for instance in [15].

The results are very satisfying for objects which do not contain pseudo random structure, as for instance tree and bush images.

Taking into account distance and rotation angle coherence significantly increases the power of the method, making the method applicable to a large number of objects.

Our work can be extended in several directions. First, in the current model, we assumed that all model patches have the same *a priori* probability to be found in the scene. We hope to increase the robustness of the proposed method by defining a prior on the patch distribution, which should make the method more efficient. We also plan to test our method with other descriptors like kAS [6].

Acknowledgments

We thank Jerome Revaud for providing us with the source code of the Zernike method described in [13] and his kind help in setting it up.

References

[1] T. F. Ansary, M. Daoudi, and J.-P. Vandeborre. A bayesian 3-d search engine using adaptive views clustering. *IEEE Transactions on Multimedia*, 9(1):78–88, 2007.

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(4):509–522, 2002.

[3] M. Brown, R. Szeliski, and S. A. J. Winder. Multi-image matching using multi-scale oriented patches. In *CVPR (1)*, pages 510–517, 2005.

[4] A. Choksuriwong, H. Laurent, C. Rosenberger, and C. Maaoui. Object recognition using local characterisation and zernike moments. In *ACIVS*, pages 108–115, 2005.

[5] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR (2)*, pages 264–271, 2003.

[6] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid. Groups of adjacent contour segments for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(1):36–51, 2008.

[7] D. Hoiem, C. Rother, and J. Winn. 3d layoutcrf for multi-view object class recognition and segmentation. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, 2007.

[8] S. Hou and K. Ramani. Calligraphic interfaces: Classifier combination for sketch-based 3d part retrieval. *Comput. Graph.*, 31(4):598–609, 2007.

[9] S.-H. Kim, I.-S. Kweon, and I.-C. Kim. Probabilistic model-based object recognition using local zernike moments. In *MVA*, pages 334–337, 2002.

[10] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, 2004.

[11] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, 1999.

[12] A. Opelt, A. Pinz, and A. Zisserman. Incremental learning of object detectors using a visual shape alphabet. In *CVPR (1)*, pages 3–10, 2006.

[13] J. Revaud, G. Lavoué, and A. Baskurt. Improving zernike moments comparison for optimal similarity and rotation angle retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):627–636, 2009.

[14] J. Shotton, A. Blake, and R. Cipolla. Contour-based learning for object detection. In *ICCV*, pages 503–510, 2005.

[15] T.S.Caetano, T. Caelli, D. Schuurmans, and D. Barone. Graphical models and point pattern matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1646–1663, 2006.

[16] H. Wang and E. R. Hancock. A kernel view of spectral point pattern matching. In *Joint Int'l Workshops Syntactical and Structural Pattern Recognition and Statistical Pattern Recognition (SSPR-SPR)*, pages 361–369, 2004.

[17] C. Wolf, J. Jolion, W. Kropatsch, and H. Bischof. Content Based Image Retrieval using Interest Points and Texture Features. In *Proceedings of the International Conference on Pattern Recognition*, volume 4, pages 234–237, 2000.

[18] W. Zhang and J. Kosecka. Generalized ransac framework for relaxed correspondence problems. In *3DPVT*, pages 854–860, 2006.