Fast and Robust Feature-based Recognition of Multiple Objects

Kai Welke, Pedram Azad, Rüdiger Dillmann
IAIM, Institute of Computer Science and Engineering (CSE)
University of Karlsruhe (TH)
P.O. Box 6980, 76128 Karlsruhe, Germany
Email: {welke, azad, dillmann}@ira.uka.de

***Abstract*— In robotics, one crucial requirement to a visual system is robust and efficient recognition of multiple objects. While in many available systems the focus is on tracking, the main problem still is to recognize objects in an arbitrary scene within a database of multiple objects. For any tracking system, recognition is needed for initialization and therefore always built in. However, the task of recognition becomes considerably harder, when learning and recognizing multiple objects. In this paper, we present a system, which accomplishes this task for textured objects robustly and efficiently. Our system is based on texture features, combining principal component analysis, k-means clustering and kd-tree search with best-bin-first strategy. We evaluated our system in several real-word scenarios, and present experimental results in a kitchen environment. Within a database of 20 objects, our system can analyze an arbitrary scene in less than 350 ms on a 3 GHz CPU.**

I. Introduction

For humanoid robotics, object recognition is an essential task. The robot has to be given the ability to identify objects in his environment, before higher level perceptional tasks or manipulation can be executed. Although object recognition has many different areas of application, like industrial image processing, augmented reality or analysis of aerial panoramas, we will introduce an approach which allows to exploit the constraints given when performing recognition on a humanoid. In this paper we focus on recognition of multiple objects in reasonable time, to allow execution in runtime on the robot.

The majority of the various approaches observed in recent object recognition systems can be categorized in two classes: model-based approaches and appearance-based approaches.

In model-based recognition systems, images of the objects as well as 3D object models are known during recognition. The task of recognizing comprises matching of the images and the object model with the actual scene. Fua et al. propose a system based on the model-based approach which allows recognizing one object in close to real-time [1]. Model-based systems allow better results for non-planar objects, however object models are hard to acquire, thus complicate the learning process.

In the appearance-based approaches, only images are used to retrieve a representation for the objects; no 3D models have to be acquired. In a fundamental paper, Nayar et al. propose a system using appearance-based methods to recognize up to 100 different objects [2]. However, the objects have to be segmented from the current scene to match image parts to objects and the task of segmenting arbitrary objects is still unsolved. To overcome the problem of segmentation, recognition systems based on local features have been proposed. In these approaches, instead of matching the whole appearance with the learned representations, significant positions in the appearance of the object are identified and the local environments of these points are learned. In the recognition phase, local features are extracted from the actual scene and matched with the learned objects features. This allows recognition without segmentation and also improves the behaviour when recognizing partially occluded objects. Different techniques have been proposed for detecting local features. Lowe proposes an approach based on Difference-of-Gaussian points to identify local features [3]. Other systems use wavelet based salient points [4], complex moments [5], Harris feature points or Hessian-Laplace feature points. Schiele et al. provide a short overview of standard techniques in [6]. While the system proposed by Schiele and Lowe identify maxima in scale-space to achieve invariance to object size, we abstain from using a scale space implementation. The stereo camera setup on a humanoid robot allows to extract feature points in both camera images, size invariance can be achieved by retrieving the 3D position of the feature point and normalizing the patch. The implementation on a stereo camera setup helps in reducing the number of patches that have to be stored per object.

Different descriptors for the local environment of detected feature points have been proposed in the literature. The scale invariant feature transform (SIFT) as proposed by Lowe et al. uses a histogram of intensities and directions of the gradients to describe the local environment. Other approaches apply a set of complex Gabor filters to the patch to retrieve a descriptor [7]. Another popular approach uses Principal Component Analysis (PCA) on the patch, to describe the local environment [8].

The system proposed in this paper uses an appearance-based approach in combination with local features points. We identify local features following the approach proposed by Shi et al. [9]. The local environment of the feature points is compressed using PCA. Contrasting the SIFT method, feature point identification and description of local features are decoupled to allow the replacement of the feature point identification method for different tasks. A feature clustering

approach is used to achieve efficient recognition of multiple objects. In related work, the benefit of feature clustering for object classification is analyzed [6]. Our approach uses clustering to reduce the search space and allow efficient recognition solely. The centroids of the identified cluster are stored in a kd-tree [10] which is traversed using an improved search strategy as proposed by Lowe et al. [11] during recognition. The resulting hypothesis is verified through Hough voting.

II. System Description

A. Overview

The task of appearance-based object recognition can be split in two major phases: the *learning phase* and the *recognition phase*. The recognition system learns the appearances of the objects during the learning phase. Views of the objects that the system is to recognize are recorded with a camera. Based on texture features extracted from the camera images, the system generates an internal representation for these objects, called *object database*. To allow robust recognition, the representations have to be invariant to changes in illumination and to transformation of the objects. Furthermore, the structure of the object database has to allow fast recognition. Figure 1 shows the process flow of the learning phase in the proposed system. The single steps are explained in the following.

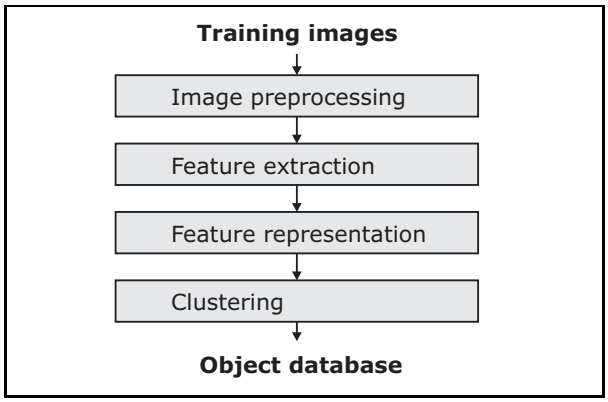


Fig. 1. Overview of the learning phase

In the preprocessing step, the color image of every view is converted into its corresponding intensity image. During the feature extraction step, significant positions in the texture of the objects are localized. A squared area around every identified position, called *patch*, is used to describe the local environment of the texture. The set of patches for an object describes the appearance of this object. During the feature representation step, the system calculates representations of these patches, which are suitable for recognition. To achieve invariance to transformation, all extracted patches of an object are warped in a given interval, thus generating a set of warped patches per identified feature point. The signal energy of all patches is normalized to achieve invariance to constant illumination changes. To allow fast recognition, the comparison between representations of the patches has to be efficient. Therefore, the intensity vectors of all warped patches are used as input for a PCA-algorithm. The PCA-compressed representations are stored in the object database along with the vectors pointing from the feature position to the center of the object. To improve recognition time, all calculated representations are classified with a standard k-means clustering algorithm and the centers of the k-means classes are stored in the object database.

In the recognition phase the calculated classes and the associated representations are used to identify the previously learned objects in arbitrary scenes. The system tries to build correspondences between features extracted from the current input scene and learned representations stored in the object database. Figure 2 shows the process flow of the recognition phase.

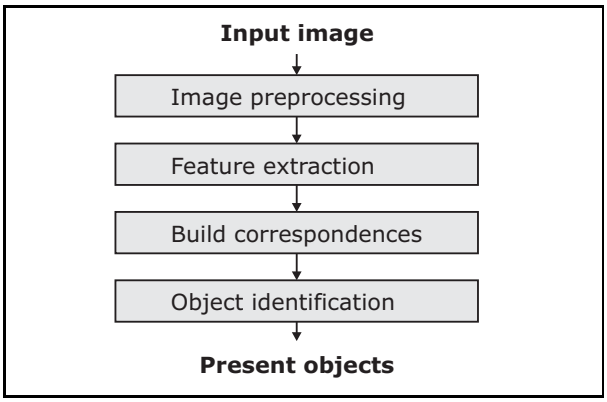


Fig. 2. Overview of the recognition phase

As in the learning phase, the image of the current scene is converted to its intensities during the preprocessing step. The resulting intensity image is used as input to the feature extraction step. The system extracts patches at significant texture positions from the whole scene. In order to build correspondences between the patches extracted from the scene and the representations stored in the database, a representation for the extracted patches has to be found which is suitable for searching the object database. During the feature representation step, the signal energy is normalized to allow invariance to constant changes in illumination. The intensity vectors of the resulting patches have to be transformed into the same vector space as the stored representations. Therefore, the vectors are transformed into the reduced PCA-eigenbase to retrieve comparable representations for the extracted patches. The system then determines, whether a known object is present in the current scene or not. For each extracted representation from the scene, a corresponding class of representations is identified in the object database and a correspondence is generated for every representation from the identified class. In general, the set of generated correspondences contains incorrect correspondences that link an unknown or falsified patch from the scene with an arbitrary database representation. In order to eliminate these incorrect correspondences, an extended Hough transform [12] is used to find bursts of correspondences that indicate the same object location. All correspondences that imply another

location of the object are discarded. An object is identified in the scene if the activation of the bin with maximum activation in Hough space is above a given threshold.

B. Feature extraction

In the feature extraction step, the system identifies locations in the designated area of the image, where the gradient has an extremum and where the curvature of the gradient is high, following the method proposed by Shi and Tomasi [9]. The parameters for the feature extraction step in the learning phase differ from those in the recognition phase. In the learning phase, the system identifies about 100–150 significant locations in the appearance of each object with the chosen parameters. In the recognition phase the parameters are adjusted to identify about 400 locations in the current scene, so that enough features are available for the correspondence building step. To describe the environment of the texture in the area of those significant locations, the locations are used as centers for squared cuts. The resulting squared image parts are called patches. The set of patches extracted from an object describes its appearance.

C. Feature representation

During the feature representation step, a description for the extracted patches has to be found, that meets three important requirements: Most importantly, the description has to be invariant to rotation and scaling of the object and to different lighting conditions to allow robust recognition. Furthermore, the structure of the description has to allow fast comparison of the descriptions in the recognition phase.

To achieve invariance to rotation and scaling, a set of warped images is generated for each extracted patch, following the equation proposed by Fua et al. [13]:

$$\begin{aligned} (\vec{n} - \vec{n_0}) &= A(\vec{m} - \vec{m_0}) - \vec{t} \\ &\text{with} \\ A &= R_\theta {R_\phi}^{-1} S R_\phi \end{aligned}$$

The rotation matrix R_θ describes the rotation of the patch in the image plane, $\vec{t}$ describes the translation of the origin. The scaling matrix S is used to scale the patch in the interval given by the two scaling values s_1 and s_2. In combination with the perspective rotation matrix R_ϕ, the scaling values s_1, s_2 are used to approximate rotations of the object outside the image plane. For this approximation, parallel projection and local planarity of the object are assumed.

To achieve invariance to illumination, the signal energy of the intensity vector of each warped patch is normalized as proposed by Nayar et al. [14]:

$$E = \sum_{n=1}^{k^2} I^2(n) = 1 \tag{1}$$

One can show, that this normalization achieves invariance to constant illumination changes c,where $I_c(n) = cI(n)$.

After warping and image normalization, the object representations consist of sets of normalized warpings for each extracted patch. The patch size in the proposed system is set to 32×32 pixels, thus the resulting normalized intensity vectors consist of 1024 dimensions each. To allow efficient comparison between these vectors, a more compact description has to be found. To reduce the number of dimensions of the intensity vectors with minimal loss in the capability to seperate them, the PCA-algorithm is deployed. All normalized intensity vectors form the input data for the PCA. The vectors are then transformed into the resulting eigenspace. Only the values corresponding to components with the twenty highest eigenvalues are kept. The reduced eigenspace values are stored in the object database.

D. Clustering

In the recognition process, the system has to identify correspondences between representations stored in the object database and representations that have been extracted from the current scene. One possibility to generate these correpondances consists in identifying the nearest neighbour for each extracted representation in the object database. The noise in the camera image and non-constant illumination changes can falsify the representations extracted from the scene. In many cases, the nearest neighbour of a falsified representation is not the correct correspondence for this representation. To increase the probability that the correct correspondence is returned and to improve computational efficiency, the database representations are sectioned into classes in the learning phase with a standard k-means clustering algorithm; the centroids of the k-means classes are stored in the object database along with the representations. During the recognition phase, the system determines the nearest centroid and returns a correspondences for each member of the class. This method improves the possibility of generating the correct correspondence, but also produces incorrect correspondences which have to be eliminated during the verification step.

Another benefit of feature clustering is the reduction of the search space. In the system an average class size of 20 representations is used. This reduces the size of the search space by a factor of 20, because only the k-means centroids have to be searched, instead of all representations.

E. Building correspondences

The search for correspondences is the most time consuming task in object recognition, even with the benefit of feature clustering. To improve the recognition time of the system, a kd-tree [10] is used to store the k-means centroids. It has been shown, that kd-trees allow efficient search for the nearest neighbour for values with twelve dimensions or less. To achieve efficient results with the 20-dimensional representations in the proposed system, the standard kd-tree algorithm has been extended with strategies for the search in high-dimensional data.

To allow searching kd-trees in constant time, David G. Lowe introduces the Lowe heuristic [15]. The Lowe heuristic value H_L specifies the maximum number of leaves, that are visited during a nearest neighbour search. The best nearest neighbour known is returned after the last leaf has been visited. However,

a search with Lowe heurstic does not always return the correct neighbour.

To increase the probability that correct neighbours are found, Lowe proposes the best-bin-first search strategy [15]. This strategy stores visited nodes along with their distances to the inquired representation in a node list while descending the tree. If a leaf is reached and the tree has to be ascended again, the strategy doesn't visit the next node according to the topology of the tree, but the node with minimal distance to the inquired representation from the list of visited nodes.

The proposed system uses a combination of Lowe heuristics with $H_L = 200$ and BBF-strategy to retrieve correct values for most inquired representations and good approximations for all other inqueries in constant time. A correspondence is generated for each representation associated with the nearest class as identified with the proposed kd-tree search.

F. Object identification

During the correspondence building step, correct as well as incorrect correspondences are generated. To decide, whether a known object is present in the scene or not, as many incorrect correspondences as possible have to be eliminated.

To verify the correctness of a correspondence, an extended Hough transform is used. The parameter space of the Hough transform comprises the x- and y-coordinate of the object center, the rotation in the plane θ and the scaling of the object that are implied by the correspondence. A different Hough space is used for every object in the database. All generated correspondences vote for the implied bin in the Hough space of the object they belong to. The amount of activiation that is issued by the vote of a correspondance depends on the size of the class it belongs to. The following rule is used for the amount of activiation: $a = \frac{1}{n_c}$.

The amount of activation a depends on the class size n_c. This rule takes into account, that representations that belong to big classes carry much less information about the presence of an object than representations belonging to small classes.

After all correspondences have voted, the bin with maximum activation is determined for every Hough space. All correspondences that activate the maximum bin are accepted. If the activation of the maximum bin is above a given threshold, the object is assumed to be present in the scene.

III. Experimental Results

A. Test scenario

We evaluated the performance of the proposed system in a kitchen scenario. The objects were learned in a kitchen cabinet and in the refrigerator. Especially in the case of recognition in the refrigerator, partial occlusion has to be handled. In our experiments we did not exploit the availabilty of the second camera. The objects were learned and recognized based on images recorded with only one camera. During warping, also transformations corresponding to translation in depth were generated. This allows us to evaluate the approach with a simple setup, the search space size can be reduced by adding support for the second camera and generating 3D feature coordinates.

Fig. 3. All objects trained during the experiment

All objects used in the experiments had to fulfil three requirements: First, enough features have to be extracted of every learned object. In the experiments, each object produced about 100–150 distinct features. Second, the extracted features have to be separable. If too many features of an object look alike, robust recognition is not possible with the proposed system. When learning objects with a reflective surface, the extracted features most likely don't describe the object itself, but the environment that is reflected on the objects surface. These objects can not be detected properly with the proposed system.

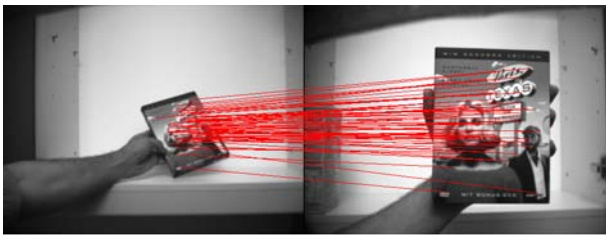
Databases with 3, 5, 10 and 20 objects were learned for the experiments. Figure 3 shows all training images used to build the object database. The parameters for perspective warping were chosen as follows:

$$\begin{aligned} \phi &\in [-25°; 25°] \\ t &\in [-2; 2] \\ \theta &\in [-45°; 45°] \\ s_1 &\in [0{,}4; 0{,}8] \\ v &\in [\frac{1}{1{,}2}; 1{,}2], s_2 = v \cdot s_1 \end{aligned}$$

This selection of parameters allows recognition of objects with rotations in the cameras image plane between -25 and 25 degrees and rotations outside the plane between -33 and 33 degrees. In addition, objects are recognized in the range of 0.8 to 0.4 of their learned sizes, thus allowing an accordingly range of object scalings.

B. Robustness

The robustness of the system can be determined by testing the invariance of the system to illumination changes, trans-



(a)

(b)

Fig. 4. Identified correspondences between learned object (right) and current scene (left). This figures show recognition with a transformed object (a) and with a partially occluded object (b).

Fig. 5. Recognition result with difficult lighting conditions

formation, and occlusion. Figure 4a shows the recognition of an object rotated in and outside the plane. The tests prove, that enough correspondences can be identified, if the object is rotated and scaled in the learned intervals. Figure 4b shows the recognition of an occluded object. Objects are recognized, if enough significant features can be extracted. In order to test invariance to changes in illumination, the recognition was performed under different lighting conditions and in different scenarios. The system has been proven to be robust under constant constant changes of illumination Figure 5 shows that the object can be recognized even in case of reflections and specular lighting, as long as enough significant features of the object can be extracted.

The overall robustness of the system was tested by recording a sequence of 300 images of three arbitrary objects from the whole set. The objects were moved manually in the learned range during the recording. Images with large motion blur were removed from the set of test images. The system did recognize all objects in the remaining frames. No false positive detections were encountered.

C. Performance

The performance of the recognition phase in the proposed system still depends on the number of representations in the object database. Though with the representations grouped with k-means clustering, the linearity factor of the dependance between the number of representations and the runtime of the recognition phase could be reduced to a minimum. Figure 6a shows the runtime in relation to the number of objects learned on an Intel Pentium 4 system with a 3 GHz CPU. Each object produced about 20000 representations, the k-means algorithm clustered the representations into about 1000 classes per object.

During the learning phase, the covariance matrix of all representations has to be calculated to receive the PCA eigenspace. In combination with the time consuming k-means clustering, the runtime of the learning phase increases from 31 minutes for three objects to 1363 minutes for 20 objects. Figure 6b shows the runtime of the learning phase in relation to the number of learned objects

IV. Conclusion

This paper proposes a new approach of fast and robust recognition of multiple objects which is well suited for the application in a humanoid robot.

The learning phase is designed very simple and can be automated almost completely. Currently, one has to obtain one image per object and has to select the area manually, in which features for this object shall be extracted. The remaining part of the learning phase is fully automated. However, the runtime of the learning phase is very high, because the covariance matrix has to be calculated for all warped patches of all extracted features and because of the time consuming k-means clustering step.

The proposed recognition scheme incorporates several invariances to achieve robustness. Invariance to transformation is achieved by learning the warped patches around significant features in the designated intervals. The amount of representations stored for the warpings can be further reduced by using a stereo camera pair and extraction 3D feature positions. The signal energy is normalized in order to achieve invariance to constant illumination changes. Due to the choice of an approach based on local features, objects can be detected even if they are partially occluded. Together with the proposed k-means clustering these invariances allow robust recognition in arbitraty scenes even under difficult conditions.

The combination of principle component analysis, the heuristic kd-tree search, the k-means clustering and the proposed extended Hough transform allows robust recognition of multiple objects without significant increase in runtime.

There are still some enhancements to the system, that we want to integrate in the future. In order to improve the recognition capabilities of the system, we want to deploy the Mahalanobis distance instead of the Euclidian distance to measure the similarity of two representations. It has been shown, that the Mahalanobis distance is well suited to build

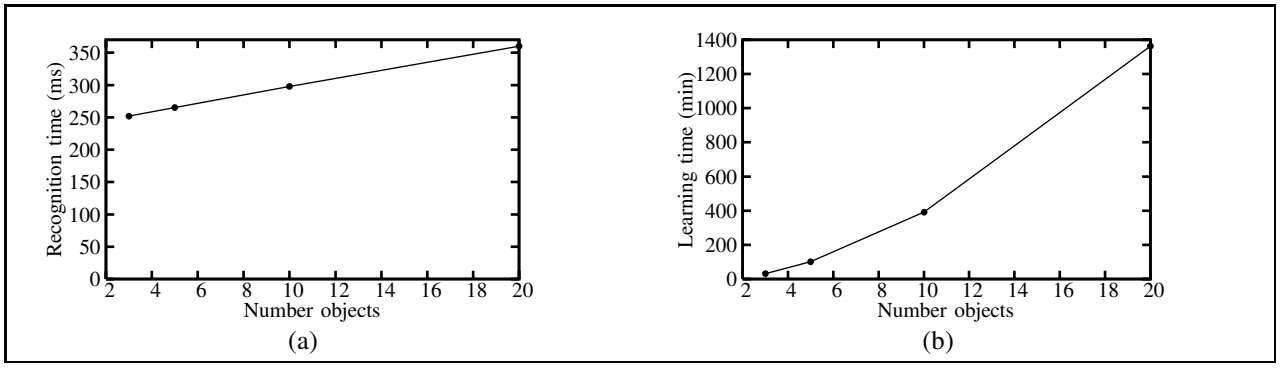


Fig. 6. Runtime of recognition and learning phase in relation to the number of objects in the database

correspondences in PCA eigenspace [16]. Also some improvements to reduce the runtime of the system will be implemented. The proposed kd-tree will be reimplemented using fixed point arithemtics, the Hough transform can be optimized by using Hash tables instead of multi dimensional arrays.

The proposed system offers all information needed for pose estimation. We will implement a pose estimation method based on the POSIT-algorithm [17], following the RANSAC paradigm [18]. First experiments have proven that the correspondences generated by the proposed system are well suited for this method.

ACKNOWLEDGMENT

The work described in this paper was partially conducted within the German Humanoid Research project SFB588 funded by the German Research Foundation (DFG) and the EU Cognitive Systems project PACO-PLUS (FP6-2004-IST-4-027657) funded by the European Commission.

REFERENCES

[1] V. Lepetit, L. Vacchetti, D. Thalmann, and P. Fua, "Fully automated and stable registration for augmented reality applications," in *Proceedings of the Second IEEE and ACM Symposium on Mixed and Augmented Reality*, 2003, pp. 93 – 102.

[2] S. Nayar, S. Nene, and H. Murase, "Real-time 100 object recognition system," in *Proceedings of the IEEE Conference on Robotics and Automation*, vol. 3, 1996, pp. 2321 – 2325.

[3] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1999, pp. 1150 – 1157.

[4] S. B. E. Loupias, N. Sebe and J. Jolion, "Wavelet-based salient points for image retrieval," in *Proceedings International Conference on Image Processing*, vol. 2, 2000, pp. 518 – 521.

[5] S.-H. Kim, I.-C. Kim, and I.-S. Kweon, "Probabilistic model-based object recognition using local zernike moments," in *IAPR workshop on Machine Vision Applications*, Nara , Japan, Dec. 2002.

[6] K. Mikolajczyk, B. Leibe, , and B. Schiele, "Local features for object class recognition," in *Proceedings International Conference on Computer Vision*, 2005, pp. 1792 – 1799.

[7] E. Murphy-Chutorian and J. Triesch, "Shared features for scalable appearance-based object recognition," in *Proceedings of the IEEE Workshop on Applications of Computer Vision*, 2005.

[8] G. Dunteman, *Principal Component Analysis*. Sage Publications, 1989.

[9] J. Shi and C. Tomasi, "Good features to track," in *Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition*, 1994, pp. 593 – 600.

[10] J. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of ACM*, vol. 18, no. 9, pp. 509 – 517, 1975.

[11] D. Lowe and J. Beis, "Shape indexing using approximate nearest-neighbour search in high-dimensional spaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1997, pp. 1000 – 1006.

[12] D. Ballard, "Generalizing the hough transform to detect arbitrary shapes," *Pattern Recognition*, vol. 13, no. 2, pp. 111 – 122, 1981.

[13] V. Lepetit, J. Pilet, and P. Fua, "Point matching as a classifictaion problem for fast and robust object pose estimastion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2004, pp. 244 – 250.

[14] H. Murase and S. Nayar, "Learning and recognition of 3D object from appearance," *International Journal of Computer Vision*, vol. 14, no. 1, pp. 5 – 24, 1995.

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91 – 110, 2004.

[16] W. S. Yambor, "Analysis of pca-based and fisher discriminant-based image recognition algorithms," Colorado State University, Tech. Rep. CS-00-103, 2000.

[17] D. DeMenthon, L. Davis, and D. Oberkampf, "Iterative pose estimation using coplanar points," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1993, pp. 626 – 627.

[18] M. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of ACM*, vol. 24, no. 6, pp. 381 – 395, 1981.