Compatibilities for the Perception-Action Cycle

Josef Pauli and Gerald Sommer

Christian-Albrechts-Universität zu Kiel
Institut für Informatik und Praktische Mathematik
Preußerstraße 1–9, D-24105 Kiel, Germany
www.ks.informatik.uni-kiel.de/~jpa

Abstract. We apply an eye-on-hand Robot Vision system for treating the following three tasks: (a) Tracking objects for obstacle avoidance; (b) Arranging certain viewing conditions; (c) Acquiring an object recognition function. The novelty is the use of so-called compatibilities between motion features and view sequence features. Under real image formation, compatibilities are more general and appropriate than exact invariants. We demonstrate the usefulness for constraining the search for corresponding features, for parameterizing correlation matching procedures, and for fine-tuning approximations of appearance manifolds.

1 Introduction

During the late eighties Computer Vision scientists realized that the human intelligence underlying the perception of the environment is not only based on views but also on accompanying actions. Since then, cameras have been mounted on agile devices in order to enable active viewing and study vision in combination with actions. Although this new paradigm of Robot Vision (or Active Vision) produced exciting solutions for problems which are too difficult for static vision, the potential usefulness is far from being fully realized [1].

Our work demonstrates the usefulness of controlled camera movements for three exemplary applications, i.e. tracking objects for obstacle avoidance, arranging certain viewing conditions, acquiring an object recognition function. In this context the theoretical concept of *invariance* is relaxed into the practical concept of *compatibility*. Regarding this, the first attempt has been undertaken by Binford and Levitt [3], who introduced *quasi-invariance* under transformations of geometric features. Our compatibility concept considers more general transformations, maybe with different types of features prior and after the mapping, and considers robot actions as the source of the transformations, and thus integrates real-world actions and perception.

We focus on compatibilities between 3D motion features and 2D view sequence features. Based on visual demonstration, statistical measurements are taken to evaluate the deviation from the exact invariance and thus specify the compatibility, which can be used in subsequent online applications. We study compatibilities for typical sub-tasks of the mentioned applications, i.e. constraining the search for corresponding features (section 2), parameterizing correlation matching procedures (section 3), and fine-tuning approximations of appearance

manifolds (section 4). For the applications we used a 6-DOF robot arm (Stäubli-RX90) and a monochrome video camera mounted on the back of the robot hand. Within a working space of a cube with sidelength $500mm$ the camera can be arranged in any position and orientation.

2 Constraining the search for corresponding features

We would like to acquire depth features from a collection of objects, e.g. bottles and cans in a refrigerator. For this purpose the camera will be translated continually in front of the objects. Gray value corners can be extracted (e.g. with SUSAN [7]) and must be tracked along the image sequence. Based on correspondences, shape-from-motion strategies can be applied to obtain the relevant information. For example, Figure 1 shows two consecutive images (left and middle) with gray value corners extracted by SUSAN, and the right image depicts motion vectors at these points. We are interested to restrict the search for corresponding corners, i.e. determine an individual disparity range for each corner.

Fig. 1. (a) and (b) Two consecutive images with gray value corners: (c) Image with motion vectors at the gray value corners.

In an experimentation (offline) phase we put a calibration pattern onto the ground plane. It depicts a regular distributed set of black dots. Both at nearest and farthest distance to the ground (i.e. the top and bottom borders of the viewing space), the camera makes a certain step of movement, respectively. Motion vectors for the calibration dots are determined in the images, resulting in two vector fields V^1 and V^2. Figure 2 shows images of the calibration pattern prior and after lateral camera translation (at top border of viewing space). The flow of dots from left to right results in vector field V^1 (not depicted). Figure 3 just shows the lengths of motion vectors of V^1, which are not constant due to large image distortions (caused by a lense with small focal length, $4mm$). For the specified camera movement, the two vector fields impose *expectations on motion vectors* which can be used later on during the online phase. Let us assume an image point p_i which originates from an arbitrary 3D point within the viewing space, and assume a step of camera movement as specified according to the calibration phase. For the image point the *angles of the motion vector* taken from V^1 or V^2 are approximately the same. Furthermore, the *length of the motion vector* must be in the interval of the relevant lengths given in V^1 and V^2. Consequently, a point p_i in the first image and a point q_j in the second image is a candidate pair for correspondence, only if the following constraints hold:

$$\Phi(V^1(p_i)) \approx \Phi(q_j - p_i) \approx \Phi(V^2(p_i)) \quad (1)$$

$$L(V^1(p_i)) \geq \|q_j - p_i\| \geq L(V^2(p_i)) \quad (2)$$

Symbol Φ denotes the angle and L the length of a vector. Just these carefully selected candidate pairs are taken for applying normalized cross correlation in order to determine the most appropriate one, as shown in Figure 1 (right image). The compatibility is represented by the two equations (1) and (2).

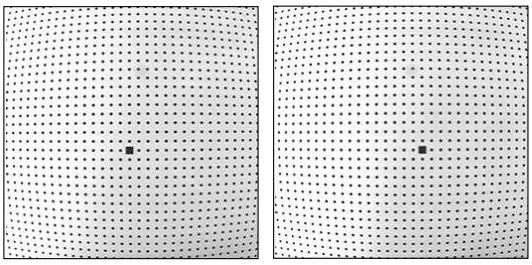

Fig. 2. Calibration pattern prior/after lateral camera translation, flow of dots from left to right.

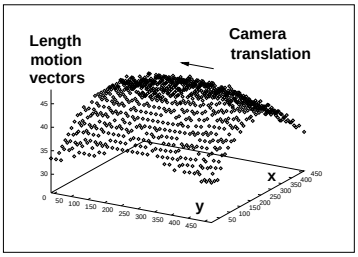

Fig. 3. Lengths of mot. vectors for lateral camera translation.

3 Parameterizing correlation matching procedures

The robot hand including the hand-mounted camera should be arranged in a certain relation to the object. This can be regarded as a sub-task of a grasping process or a sub-task leading to optimal viewing of an object. A servoing mechanism will be applied which does the arrangement step by step and is based on continual visual feedback and correlation matching in the series of images. In section 2 we treated exemplary the case of camera translation, and now we consider compatibilities for the case of camera rotation. If a camera is rotating around the optical axis, which is normal to the object surface, then *log-polar transformation (LPT)* can be applied to the gray value images [4]. The motivation is that the transformed object pattern is shifting instead of rotating, which makes the correlation matching more efficient during the servoing process. Figure 4 shows two images of an integrated circuit (IC) object under rotation by a turning angle of 90°. These are two examples from a collection of 24 images taken under angle offset of 15°, respectively. Figure 5 shows the horizontal translation of the log-polar transformed pattern of the rotating object.

However, in a view sequence perfect invariance only holds for a flat 2D object without any side faces, and a simulated pinhole camera is assumed whose optical axis must be kept normal to the object surface. In realistic applications, resampling error occur certainly, the objects are of three-dimensional shape presumably, the camera objectives may cause unexpected distortions, and possibly the optical axis is not exact normal to the object surface (misalignment). Because of these realistic imponderables, certain variations of the LPT patterns will occur. We are interested in determining the real deviations from invariance in order to obtain tolerance parameters for correlation matching.

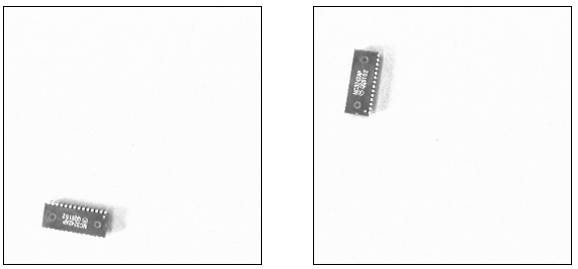

Fig. 4. Integrated circuit object under rotation by turning angle 90°.

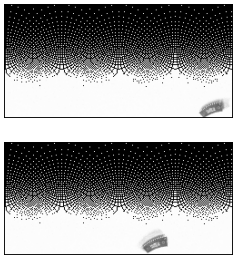

Fig. 5. Horizontal translation of LPT pattern.

By demonstrating sample objects and performing typical camera rotations relative to the objects, one can make measurements of real deviations from invariance, i.e. *actual variations of the LPT patterns*. Despite of these variations, it is expected that the manifold of LPT patterns is much more compact and easier to describe than the original manifold of appearance patterns. For example, presumably, a *single multi-dimensional Gaussian*, specified by a center vector and a certain covariance matrix, may approximate the variation.

For illustration, we perform a simple experiment which is based on histograms of edge orientations. Specifically, orientations of gray value edges are considered in order to demonstrate the influence of LPT to a rotating 3D object, i.e. measuring the deviation from pure pattern translation in the log-polar transformed image. The image library consists of 24 images, as mentioned above. The histograms should be computed from the relevant area of the LPT image containing the object pattern, respectively. To simplify this sub-task a nearly homogeneous background has been used such that it is easy to extract the gray value structure of the IC object. We compute for the extracted LPT patterns a histogram of gradient angles of the gray value edges, respectively.

Figure 6 (left) shows a histogram determined from an arbitrary image in the library. The mean histogram is computed from the LPT patterns of the whole set of 24 images, shown in Figure 6 (middle). Next, we compute for each histogram the deviation vector from the mean histogram. From the whole set of deviations once again a histogram is computed, which is shown in Figure 6 (right).

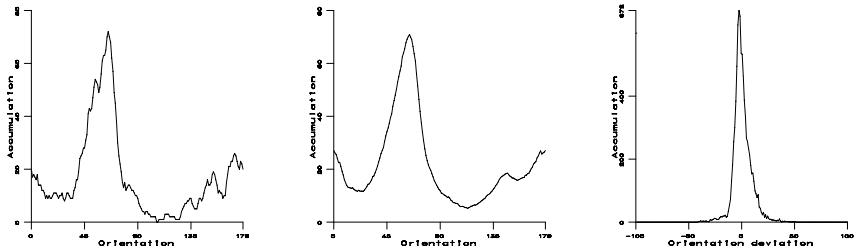

Fig. 6. (a) Histogram of edge orientations under LPT for one image; (b) Mean histogram for several images; (c) Accumulation of orientation deviations.

This latter histogram can be approximated as a Gaussian with the maximum value at 0 and the Gaussian turning point approximately at the value $\sigma = 5$. Under ideal (simulated) conditions the Gaussian would be an impulse function with extent 0. However, the real value of σ is a measure for the deviation from perfect invariance. It can be used to parameterize approaches of pattern matching, e.g. specifying thresholds for the coefficient of normalized cross correlation in order to obtain reasonable matching hypotheses.

4 Fine-tuning manifold approximations for recognition

For the recognition of a scene object in an image we need to have an appropriate recognition function. This function can hardly be implemented manually and instead should be learned automatically in the task-relevant environment. Based on a robot-controlled process of taking sample views we can incorporate action-related information for improving the generalization in the learning mechanism.

Appearance-based object recognition
A holistic learning approach can be applied which is based on 2D appearance patterns of the relevant objects or response patterns resulting from specific filter operations. The main interest is to represent or approximate the pattern manifolds such that an optimal compromise between *efficiency, invariance and discriminability* of object recognition is achieved. It is essential to keep these manifolds as simple as possible, because the complexity is correlated to the time needed for object recognition. In section 3 we restricted camera poses and movements and thereby reduced the manifold complexity by LPT. However, in this section we accept general viewing poses. Apart from the efficiency criterion the recognition function must respond with constant high values for any appearance of the object (invariance criterion), and must be able to discriminate between target and other objects (discriminability criterion).

The most popular approach of manifold approximation is based on principal component analysis (PCA) for a collection of views [5]. This is done for each object leading to an object-specific Eigenspace, respectively. An unknown view can be recognized by computing proximity values to the training samples in the Eigenspaces, and determining the most relevant manifold. An improvement of this one-nearest-neighbor approach is obtained by applying a clustering approach for the purpose of generalization. Closely located training samples are clustered and the clusters approximated as a multi-dimensional Gaussian, respectively. However, clustering procedures such as ISODATA search for neighboring elements according to simple metric, and do not consider any inherent topology between training samples. For example, if training samples are acquired by consecutively rotating the camera around the object, then we know in advance that the pattern variation can be approximated as a one-dimensional course in the space of patterns. Consequently, the clustering procedure should generate segments of this course by taking the succession of training views into account. By *imposing a topology* onto the collection of sample views, which is obtained from the process of image taking, we can cluster more adequately.

Role of temporal context in object recognition

In addition to the one-dimensional topology we also take advantage of the *temporal continuity* of gray values between the views in the image sequence.[1] For an object under rotation the temporal continuity can be observed exemplary in a series of histograms of orientations of gray value edges. Figure 7 shows four gray value images (a,b,c,d) of a transceiver box which has been rotated slightly in four discrete steps of 5°. Figure 8 depicts the overlay of four histograms of edge orientations for these four images (but suppressing the gray values of the background). The histogram curves move to the right continually under slight object rotation.[2] These sequential correlations between consecutive images hold for small changes in the relation of object and camera. They are considered for fine-tuning the manifold approximation.

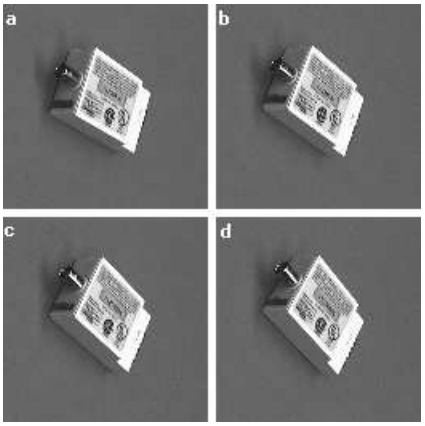

Fig. 7. Four gray value images of a transceiver box under rotation in discrete steps of turning angle 5°.

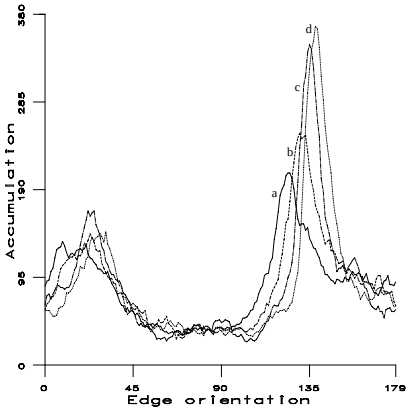

Fig. 8. Overlay of four histograms of edge orientations computed for the four images in Figure 7.

Incorporating temporal context for manifold approximation

Let us assume that the clustering is already performed under the constraint of a one-dimensional topology. This leads to a representative view for each cluster, respectively, which will be taken as *seed views* for manifold approximation. A *sequence of Gaussian basis functions* is used for approximating the one-dimensional course in the space of patterns. Each seed view is the basis for specifiying the center of a multi-dimensional Gaussian with the dimension equal to the number of pixels. Each Gaussian is almost hyper-spherical except for one direction whose Gaussian extent is stretched. The exceptional direction at the current seed view is determined on the basis of the difference vector between the previous and the next seed view. For illustrating the principle, we take two-dimensional points which represent the seed views. Figure 9 shows a series of three seed views, i.e.

[1] The importance of temporal context in object recognition is well-known [2].

[2] The variation of the accumulation values is due to changing lighting conditions or due to the appearing or disappearing of object faces.

previous, current and next seed view ($\boldsymbol{X}^s_{i-1}$, $\boldsymbol{X}^s_i$ and $\boldsymbol{X}^s_{i+1}$). At the current seed view the construction of an elongated Gaussian is depicted. Actually, an ellipse is shown which represents the contour related to a certain Gaussian altitude.

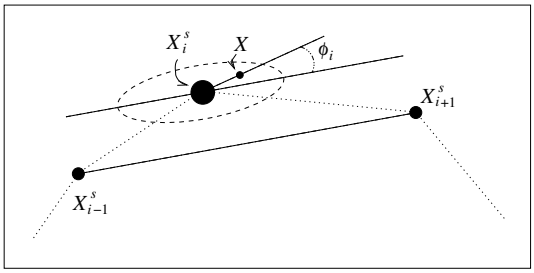

Fig. 9. Constructing hyper-ellipsoidal basis functions for time-series of seed vectors.

The Gaussian extent along this exceptional direction must be defined such that the significant variations between successive seed views are considered. For orthogonal directions the Gaussian extents are only responsible for taking random imponderables into account such as lighting variations. Consequently, the Gaussian extent along the exceptional direction must be set larger than the extent along the orthogonal directions. It is reasonable to determine the exceptional Gaussian extent dependent on the euclidean distance measurement between the previous and the next seed view. We avoid mathematical details because they are simple. However, it is worth to mention a similarity of this approach of manifold approximation with the so-called "oriented particle system" for surface modeling, introduced by Szeliski and Tonnesen [8].

Applying the fine-tuned manifold for object recognition

Although our approach is very simple, both efficiency and robustness of the recognition function increases significantly. The usefulness of constructing elongated Gaussians is illustrated for recognizing the transceiver box in Figure 7. For learning the recognition function the object is rotated in steps of 10° leading to 36 training images. All of them are used as seed images (for simplicity). The computation of gradient magnitudes followed by a thresholding procedure yields a set of gray value edges for each seed image. From each thresholded seed image a histogram of edge orientations is computed. A Gaussian basis function (GBF) network is installed by defining elongated GBFs according to the mentioned approach. Histograms of the seed images are used as the Gaussian center vectors and the Gaussians are modified based on previous and next seed histograms (and applying a user-defined stretching factor). In the GBF network the combination factors for the Gaussians are determined by the pseudo inverse technique.

For assessing the network of elongated Gaussians, we also construct a network of spherical Gaussians and compare the recognition results computed by the two GBF networks. The testing views are taken from the transceiver box but different from the training images. The testing data are subdivided in two categories. The first category consists of histograms of edge orientations arising from images with a certain *angle offset* relative to the training images. Temporal continuity of object rotation is considered purely. For these situations the

relevant recognition function has been trained particularly. The second category consists of histograms of edge orientations arising from images with *angle offset and are scaled*, additionally. The recognition function composed of elongated Gaussians should recognize histograms of the first category robustly, and should discriminate clearly the histograms of the second category. The recognition function composed of spherical Gaussians should not be able to discriminate between both categories due to an increased generalization effect, i.e. accepting not only the angle offsets but also scaling effects. The desired results are shown in the diagrams of Figure 10. By applying the recognition function of spherical Gaussians to all testing histograms, we can hardly discriminate between the two categories (left). Instead, by applying the recognition function of elongated Gaussians, we can define a threshold for discriminating between both categories (right).

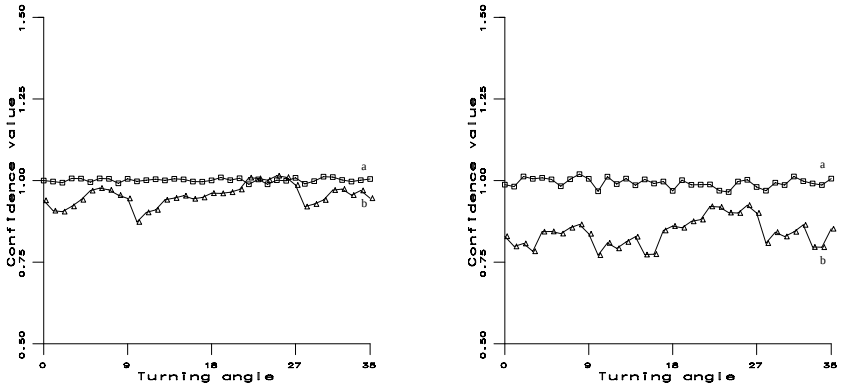

Fig. 10. Confidence values of recognizing an object based on histograms of edge orientations. For testing, the object has been rotated by an offset angle relative to the training images (result in curve a), or the object has been rotated and the image has been scaled additionally relative to the training images (result in curve b). (Left) Curves show the courses under the use of spherical Gaussians, both categories of testing data can hardly be distinguished; (Right) Curves show the courses under the use of elongated Gaussians, both categories of testing data can be distinguished clearly.

5 Summary and discussion

For an eye-on-hand system we presented three typical applications, i.e. tracking objects for obstacle avoidance, arranging certain viewing conditions, and acquiring an object recognition function. The concrete tasks have been to constrain the search for corresponding features, to parameterize correlation matching, and to fine-tune appearance manifolds. For solving the first task, specific steps of motion are performed during an experimentation phase in order to acquire constraints for motion vectors. By restricting the kind of motion to these specific ones during the online phase we can exploit the acquired constraints in the search for correspondences. For solving the second task, a specific course of motion is performed during an experimentation phase in order to determine real deviations from a theoretical invariance, i.e. the variation of an LPT pattern under

camera rotation. The distribution of deviations has been approximated by a Gaussian. The Gaussian extent can be used to determine a threshold in procedures which make use of correlation matching. For solving the third task, the camera is moved step by step for acquiring appearance patterns from an object. The pattern manifold is fine-tuned by making use of the known one-dimensional topology and the temporal continuity of the gray-values.

All three examples have in common that specific movements of the camera lead to certain changes in the images. From an abstract point of view, these are compatibilities between 3D motion features and 2D view sequence features. They are approximated during an experimentation phase based on statistical evaluations. Also, the examples show the usefulness of repeatable actions, i.e. the pre-specified actions in the experimentation phase must be repeatable in the application phase. The usefulness is due to the applicability of action-based information for supporting image processing in the application phase.

Apart form compatibilities for the perception-action cycle, which have been treated exemplary in this work, we also studied other compatibilities for the purpose of boundary extraction (published in [6]). The advantage is to reduce the amount of object-specific knowledge and instead make extensive use of constraints which are inherent in the three-dimensional nature of objects and in the process of image formation. For high-level Robot Vision applications a further category of compatibility is of interest. It is the compatibility between a deliberate plan (e.g. a strategy for solving a task) and the concrete servoing process (which is based on visual feedback). Generally, a compromise is needed between plan fulfillment and plan adjustment with the latter being triggered by requirements in the observed reality. Our approach of considering such compatibilities is based on dynamic potential fields (a publication will be prepared soon).

References

1. Aloimonos, Y., Fermüller, C.: Analyzing action representations. Workshop on Algebraic Frames for the Perception-Action Cycle, LNCS **1888** (2000) 1-21
2. Becker, S.: Implicit learning in 3D object recognition – The importance of temporal context. Neural Computation **11** (1999) 347-374
3. Binford, T., Levitt, T.: Quasi-invariants – Theory and exploitation. Image Understanding Workshop (1993) 819-829
4. Bolduc, M., Levine, M.: A review of biologically motivated space-variant data reduction models for robot vision. Comp. Vis. and Image Underst. **69** (1998) 170-184
5. Murase, H. Nayar, S.: Visual learning and recognition of 3D objects from appearance. International Journal of Computer Vision **14** (1995) 5-24
6. Pauli, J.: Compatibilities for boundary extraction, Symp. der Deutschen Arbeitsgem. für Mustererkennung, Informatik aktuell Springer-Verlag (2000) 468-475
7. Smith, S., Brady, J.: SUSAN – A new approach to low level image processing. International Journal of Computer Vision **23** (1997) 45-78
8. Szeliski, R., Tonnesen, D.: Surface modeling with oriented particle systems. ACM SIGGRAPH Computer Graphics Annual Conference (1992) 185-194