Estimating the Pose of Phicons for Human Computer Interaction

Daniela Hall* and James L. Crowley

Projet PRIMA — Lab. GRAVIR–IMAG
INRIA Rhônes–Alpes
655, avenue de l'Europe
38330 – Montbonnot Saint Martin, France

Abstract. Physical icons (phicons) are ordinary objects that can serve as user interface in an intelligent environment. This article addresses the problem of recognizing the position and orientation of such objects. Such recognition enables free manipulation of phicons in 3D space.
Local appearance techniques have recently been demonstrated for recognition and tracking of objects. Such techniques are robust to occlusions, scale and orientation changes. This paper describes results using a local appearance based approach to recognize the identity and pose of ordinary desk top objects. Among the original contributions is the use of coloured receptive fields to describe local object appearance. The view sphere of each object is sampled and used for training. An observed image is matched to one or several images of the same object of the view sphere. Among the difficult challenges are the fact that many of the neighborhoods have similar appearances over a range of view-points.
The local neighborhoods whose appearance is unique to a viewpoint can be determined from the similarity of adjacent images. Such points can be identified from similarity maps. Similarity maps provide a means to decide which points must be tested to confirm a hypothesis for correspondence matching. These maps enable the implementation of an efficient prediction–verification algorithm.
The impact of the similarity maps is demonstrated by comparing the results of the prediction–verification algorithm to the results of a voting algorithm. The ability of the algorithm to recognize the identity and pose of ordinary desk-top objects is experimentally evaluated.

Keywords: Object Recognition, Appearance-Based Vision, Phicons

1 Introduction

Phicons provide an important new mode for man machine interaction in an intelligent environment [5]. Phicons are physical objects whose manipulation can serve as a numerical interaction device. For example, an environment may be configured so that grasping an object determines a functional context for

* Daniela.Hall@inrialpes.fr

interaction, and the manner in which the object is manipulated determines input parameters. Another example would be to use a phicon as a 3D space mouse in which the position and orientation of the object determines the position and orientation of the cursor. The use of the phicon needs no explication and appears for this reason much more natural to the user. When coupled with steerable cameras, the use of computer vision to sense the manipulation of phicons can permit such input at any location within the intelligent environment. This should allow phicons to become a graspable, easy to use, human computer interaction mode.

This article describes an algorithm for pose estimation of manipulated phicons based on local appearance. Local appearance techniques are robust to occlusions[2] and to scale changes[1] and they have recently been demonstrated for the recognition of objects. During the matching process of an observed image to a set of view sphere images, many ambiguous points are present. We propose a method to determine characteristic points for a particular view, which avoid ambiguous points and enables an efficient prediction–verification algorithm for view point determination.

2 Pose recognition

The problem of pose recognition can be transformed to the problem of determining the relative position between the camera and the object. Although the illumination of an object changes in an uncontrolled environment while it is turned in space, the recognition of the view point of an object serves as the basis for the recognition of phicon manipulation.

The experiments in this article are based on an object recognition algorithm using coloured receptive fields [4]. The basis for the receptive field are Gaussian derivatives in order to profit from scalability and the possible orientation to any arbitrary direction. These two properties allow recognition independent from scale and orientation [6,3].

Coloured receptive fields can be adapted to the recognition problem by the selection of different derivatives in the luminance and chrominance channels. An example of a coloured receptive field used for this application is shown in figure 1. The receptive field is oriented vertically with standard deviation $\sigma = 2.3$. In the coloured receptive field technique the luminance channel is maintained and complemented with two channels based on chrominance. The RGB coordinate system is transformed to the Luminance/Chrominance space according to equation (1). The chrominance channels are described using colour-opponent receptive fields. Luminance is known to describe object geometric structure while chrominance is primarily used for discrimination.

$$\begin{pmatrix} l \\ c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ -\frac{1}{2} & -\frac{1}{2} & 1 \\ \frac{1}{2} & -\frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} r \\ g \\ b \end{pmatrix} \tag{1}$$

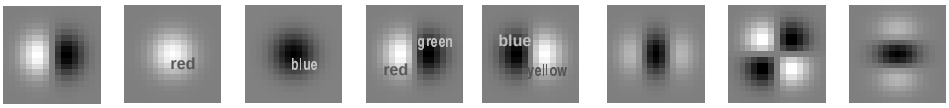

Fig. 1. Vertically oriented coloured receptive field with $\sigma = 2.3$ (1^{st} order Gaussian derivative in the luminance channel, 0^{th} and 1^{st} Gaussian derivatives in the two chrominance channels, and 2^{nd} order Gaussian derivatives in the luminance channel).

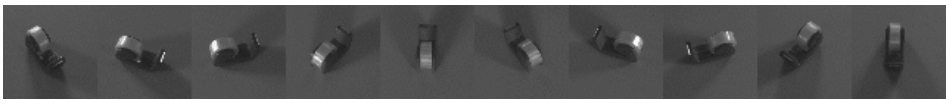

Fig. 2. Images of latitude 58° from the sampled view sphere.

During the training phase the local appearance of each point neighborhood within a training image is measured by the receptive field normalized to the local scale and oriented to the dominant direction. The receptive field response is stored in a hash table for fast access together with identification of the image and coordinates of the center of the point neighborhood. This supplemental information enables recognition of objects and their pose.

The object recognition algorithm is extended to view point recognition as follows. Images of an object from different view points serve as training base. The object is segmented from the background, in order to consider only object information. Pose is determined by recognizing which image of the view sphere most closely resembles the observed image. For this purpose the View Sphere Database[1] has been created, which contains images of a dense sampling of a geodesic view sphere for 8 objects.

3 Prediction–verification with similarity maps

An unexpected difficulty occurs when using a relatively dense sampling of the view sphere: Adjacent view-points demonstrate similar appearance. As a results, the recognition process must take into account the geometric constraint between several observed points and their correspondences in the training images. The question for the design of a good prediction–verification algorithm is: Which points lead to reliable recognition? How can these points be found?

Reliable recognition can be performed when one or several discriminant or characteristic points for a particular view are found. A characteristic point for a particular view $V_A(\varphi, \lambda)$ defined by latitude φ and longitude λ of object A is

[1] The View Sphere Database: Information and data is available at `http://www-prima.inrialpes.fr/Prima/hall/view_sphere.html`.

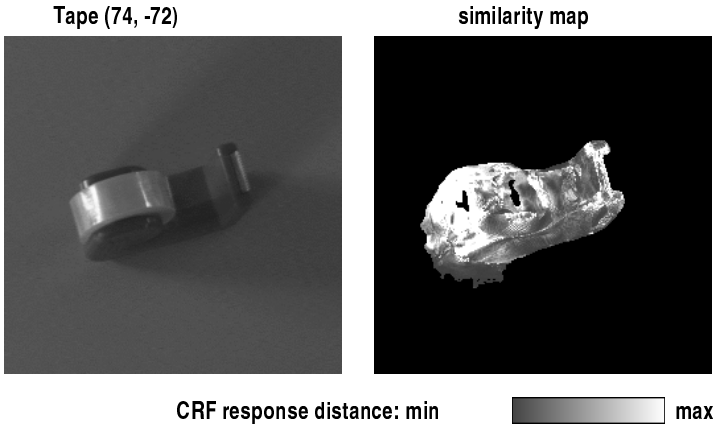

Fig. 3. Image and its similarity map. The map is computed by measuring the similarity of corresponding points between the image and its neighboring images. The images are pre-segmented and the background is not considered.

a point that is different in $V_A(\varphi, \lambda)$ than in all other views and allows to distinguishes $V_A(\varphi, \lambda)$ from other views. The most ambiguities and confusions during a matching process occur between neighboring views. A useful approximation for the determination of characteristic points for $V_A(\varphi, \lambda)$ is to restrict the search of characteristic points of $V_A(\varphi, \lambda)$ to neighboring view images $V_A(\varphi_k, \lambda_k)$.

Receptive fields provide a description of point neighborhoods. The distance of the receptive field responses in the descriptor space is a measure for the similarity of point neighborhoods. This similarity property is used to determine a similarity map for all training images which indicates if a particular point is characteristic for this training image or not (see figure 3).

To determine the similarity map of an image $V_A(\varphi, \lambda)$, a structure is trained with all neighbor images of $V_A(\varphi, \lambda)$. Then point correspondences are searched between the points in $V_A(\varphi, \lambda)$ and neighboring views $V_A(\varphi_k, \lambda_k)$. This is done by simply searching the closest receptive field response to the observed response within the trained structure. For each point, the distance of the observed receptive field response and the trained receptive field response is measured. These values form the similarity map. Small values signify high similarity of a particular point within the neighboring images. Such a point is ambiguous and can not serve to identify the pose. Whereas higher values signify that no similar correspondence has been found in the neighboring images. Such points are characteristic and can serve to distinguish the view $V_A(\varphi, \lambda)$ from its neighboring views.

Let $V_A(\varphi_0, \lambda_0)$ be the image of object A from view point with latitude φ_0 and longitude λ_0. Let $V_A(\varphi_k, \lambda_k), k = 1 \ldots u$ be the u direct neighbor images of

$V_A(\varphi_0, \lambda_0)$ of object A. Let

$$\boldsymbol{m}_{V_A(\varphi_l,\lambda_l)}(i,j) = \mathrm{CRF}(V_A(\varphi_l,\lambda_l), i, j) \tag{2}$$

be the coloured receptive field (CRF) response of image $V_A(\varphi_l, \lambda_l)$ at position (i, j). The characteristic $c_{V_A(\varphi_0,\lambda_0)}$ of a point (i, j) in image $V_A(\varphi_0, \lambda_0)$ is

$$c_{V_A(\varphi_0,\lambda_0)}(i,j) = \min_{x,y,k} \left(\| \boldsymbol{m}_{V_A(\varphi_0,\lambda_0)}(i,j) - \boldsymbol{m}_{V_A(\varphi_k,\lambda_k)}(x,y) \| \right) \tag{3}$$

$\|.\|$ is the Mahalanobis distance in CRF space taking into account the distribution of the CRF responses. The point characteristic $c_{V_A(\varphi_0,\lambda_0)}$ forms the similarity map $S_{V_A(\varphi_0,\lambda_0)}$.

$$S_{V_A(\varphi_0,\lambda_0)} = ((c_{ij}))_{i=0\ldots n, j=0\ldots m}, \text{with } c_{ij} = c_{V_A(\varphi_0,\lambda_0)}(i,j) \tag{4}$$

where $((.))$ forms an image.

The computation of the similarity maps is a method to determine characteristic points for images over the view sphere. If a characteristic point in an observed image can be found at a predicted position, the object and its pose can be reliably determined. The following prediction–verification algorithm uses the precomputed similarity maps to determine characteristic points and search them in the observed image to verify a hypothesis.

In a first step a hypothesis list for an image is obtained by applying the receptive field to a point neighborhood and searching similar receptive field responses in the structure. For each hypothesis of the hypothesis list characteristic points are searched in the hypothesis image $V_{Obj}(\varphi, \lambda)$ using the similarity map $S_{V_{Obj}(\varphi,\lambda)}$. For each selected characteristic point a region of interest in the observed image is computed, that fulfills the spatial geometric constraints required by the hypothesis image. Within this region of interest the corresponding point is searched by minimizing similarity and spatial distances. If the similarity is below a threshold, the confirmation counter of the hypothesis is incremented. The hypothesis with the highest confirmation counter is returned.

This algorithm uses several points to verify the hypothesis. Only those points are considered for testing for which the geometric constraint is fulfilled. The algorithm is fast, because the point within the region of interest is returned, whose receptive field response minimizes the distance in the descriptor space. Note that the characteristic points of the training images can not be found by any other interest point extraction technique, because the characteristic of a point depends only on the variance of features in the training images.

4 Experimental Results

Since the goal of this project is phicon recognition in an intelligent environment, the database used in the experiments contains 8 different objects from an office environment such as scissors, a tape, and a pen representing natural phicons. The view sphere is sampled according to a geodesic sphere with a spherical

distance of $\frac{\pi}{10}$ between images. This corresponds to a sampling of 90 images per half sphere. A test set of a 100 images is constructed such that a test image has exactly 3 nearest neighbors among the training images. Each image $V_{Obj}(\varphi, \lambda)$ is identified by its latitude φ and longitude λ with respect to the view sphere.

In order to illustrate the utility of the similarity maps, the results of two different pose recognition algorithms are compared. A voting algorithm serves as benchmark. The best hypothesis is determined by considering the votes of 4 direct spatial neighbor points of each hypothesis of the hypothesis list. The hypothesis with the maximum number of votes is returned. The results of the vote algorithm are then compared to the results of the prediction–verification algorithm described above. The difference between the two algorithms consists in the choice of the verification or voting points. In the voting algorithm the voting points are located north, south, east and west of the test point. Whereas in the prediction–verification algorithm, the verification points are selected according to their characteristics $c_{V_A(\varphi,\lambda)}$ obtained from the similarity map $S_{V_A(\varphi,\lambda)}$ of the hypothesis image $V_A(\varphi, \lambda)$.

The view point of the test images $V_{Obj}(\alpha, \beta)$ is recognized by matching the observed image to one or several of the training base. The result of the recognition process is evaluated by computing the spherical distance of the hypothesis view point $V_{Obj}(\varphi, \lambda)$ and the view point $V_{Obj}(\alpha, \beta)$ of the observed image. In table 1 the percentage of images whose pose is detected correctly is displayed. For each algorithm the percentage of correct pose determination with two different precisions are given. The higher precision corresponds to the percentage of images for which one of the closest neighbors is found. The resulting precision is a spherical distance of $< \frac{\pi}{16}$. The other percentage is measured with a precision of $< \frac{\pi}{9}$, which is sufficient for many applications. Lower precision results in higher reliability of the results.

Object	recognition by vote		recognition by pred-verif	
	$d < \frac{\pi}{16}$	$d < \frac{\pi}{9}$	$d < \frac{\pi}{16}$	$d < \frac{\pi}{9}$
Eraser	0.52	0.68	0.66	0.81
Pen	0.52	0.67	0.75	0.86
Scissors	0.35	0.48	0.47	0.65
Sharpener	0.46	0.53	0.67	0.80
Stapler	0.29	0.40	0.48	0.65
Tape	0.33	0.75	0.80	0.90
Protractor	0.54	0.68	0.57	0.70
Vosgienne	0.58	0.75	0.59	0.76

Table 1. Comparison of the pose recognition results of a voting algorithm and prediction–verification using similarity maps. Displayed is the percentage of images for which the correct pose has been detected with a precision of either $< \frac{\pi}{16}$ or $< \frac{\pi}{9}$.

The results confirm the superiority of the prediction–verification algorithm. The recognition rates shown in table 1 are remarkable, because the change in

view point between test and training images is up to 9 degree in latitude and up to 36 degree in longitude. The object is illuminated with 2 diffuse spot lights. In more than half of the images specularities are present. Specularities change according to the Lambertian rule and alter the appearance of the object. Both algorithms can deal with this difficulty due to the locality of the receptive fields.

5 Conclusions and Outlook

This article proposes a prediction–verification algorithm for a reliable recognition of camera view points on an object. This serves as the basis for pose recognition of phicons during manipulation in an intelligent environment. The problem of ambiguities between images of the training set is solved by computing similarity maps allowing a determination of characteristic points for a particular view and leading to a reliable recognition.

The next step will be to test the performance of this algorithm for the recognition of the manipulation of objects. The problem of partial occlusions and rapidly changing orientation in 3D space will be adressed. The use of phicons that can be freely manipulated will have an significant impact on man machine communication, since any physical object can than serve as input device.

References

1. O. Chomat, V. Colin de Verdière, D. Hall, and J.L. Crowley. Local scale selection for gaussian based description techniques. In *European Conference on Computer Vision (ECCV 2000)*, Dublin, Ireland, June 2000.
2. V. Colin de Verdière. *Représentation et Reconnaissance d'Objets par Champs Réceptifs*. PhD thesis, Institut National Polytechnique de Grenoble, France, 1999.
3. W.T. Freeman and E.H. Adelson. The design and use of steerable filters. *Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, September 1991.
4. D. Hall, V. Colin de Verdière, and J.L. Crowley. Object recognition using coloured receptive fields. In *European Conference on Computer Vision (ECCV 2000)*, Dublin, Ireland, June 2000.
5. H. Ishii and B. Ullmer. Tangible bits: Towards seamless interfaces between people, bits and atoms. In *Computer Human Interfaces (CHI '97)*, Atlanta, USA, March 1997.
6. T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.