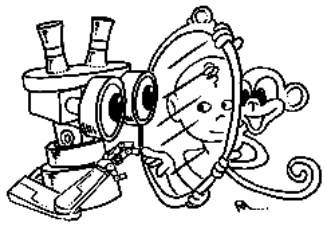


MIRROR

IST–2000-28159

Mirror Neurons based Object Recognition

Deliverable Item 2.3

Visual Primitives for Object Identification

Delivery Date: May, 2002

Classification: Internal

Responsible Person: Prof. José Santos-Victor – Instituto Superior Técnico (IST)

Partners Contributed: Alexandre Bernardino, Manuel Lopes (IST)

Short Description: This deliverable describes the visual primitives developed for being implemented in the experimental artifact. In more detail, this deliverable describes:

i) a methodology developed for computing the view point transformation between the artifact's own arm and the demonstrators when performing imitation. Even if this is a high level behavior that exceeds the scope, it also includes processes for hand/arm segmentation in video sequences.

ii) an approach for the computation of 3D dense depth maps from binocular disparity channels using log-polar images;

iii) low-level processes and software for extracting image corners and compute the normal flow from image sequences;

These visual primitives will be integrated in the final artifact for conducting experiments on learning and imitation at a later stage of the project.

Project funded by the European Community under the "Information Society Technologies" Programme (1998-2002)

Content list

1. Introduction

The visual primitives described in this deliverable have been developed for different levels of application in the project, ranging from low-level feature extraction, including medium level 3D processing and figure-ground segmentation to the visual transformations involved in imitation and the required image processing tools for segmenting one's arm/hand in an image/video sequence.

Section 2 is devoted to the problem of imitating harm/hand gestures. One of the aspects considered is that of the Visual Transformation between the self-image when looking at one's arm or that observed when looking at someone else's body. In addition it describes the visual processing designed for segmenting one's arm/hand in an image.

Section 3 details how a 3D dense depth map can be obtained by utilizing a set of disparity channels computed from log-polar images. Section 4, describes briefly the low-level feature extraction processes that have been designed.

It is of course expected that these Visual Primitives will evolve throughout the duration of the project either through adaptation to experimental needs or by including new Visual Primitives as necessary.

2. Viewpoint Transformation and hand/harm segmentation

We have studied the overall problem of endowing an artifact with the capability to imitate someone's gestures. Even if imitation is a high-level capability we describe this work in this report. The architecture is composed of two main blocks: a *Sensory-Motor Map* (SMM) and a *View-Point Transformation* (VPT) module.

The *sensory-motor map* relates images of the arm to the required forces (or arm joint angles) that produce such an image configuration of the arm. It thus involves not only the arm (inverse) kinematics but also the camera geometry. It is learned during an initial stage where the robot performs random arm movements and uses the image observations to estimate the sensory-motor transformation. Once the SMM has been learned, the artifact is in theory able to generate the necessary arm torques or joint configurations to place the arm at a certain image posture.

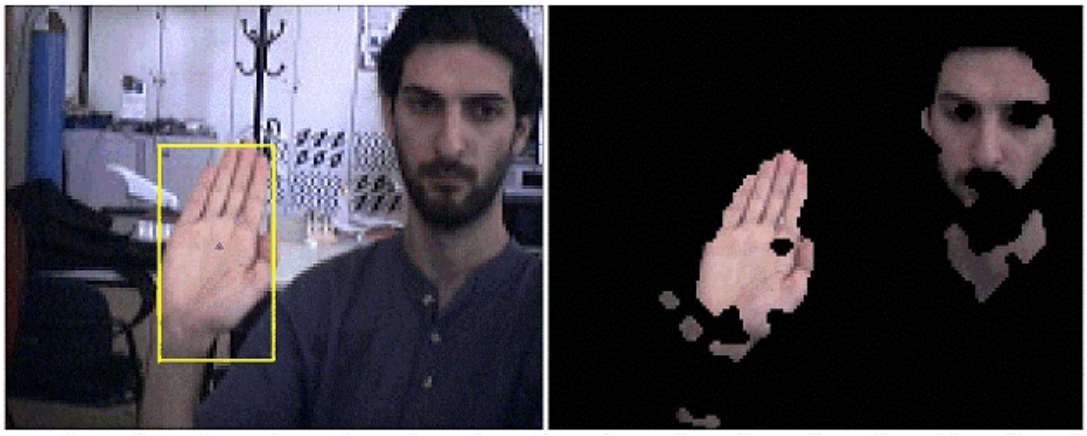
The SMM alone is not sufficient to allow the artifact to imitate the demonstrator's gestures. In fact, the demonstrator gestures are imaged in a different position when compared to the artifact's own arm, when performing the same gesture. This is illustrated in the figure on the left side, where both the arm of a teacher and that of a student perform similar gestures but they are perceived in very different image configurations, due to perspective.

There is thus the need to transform the image of the demonstrator's arm to a new configuration that corresponds to the image

that the artifact would observe if its own arm were performing that same gesture. We refer to this process as the *View Point transformation*. Different classes of VPTs can be used depending on the task. The transformation can reproduce the full 3D transformation to align the demonstrator and the artifact's bodies. However, in some other cases simpler geometric transformations suffice and this fact is vastly illustrated in experiments conducted with children or infants.

Once the SMM and VPT have been defined and implemented, the system is able to observe the gestures of a demonstrator, transform the observed image to the so-called allo-image, as if observing its own arm and finally perform the gesture. In the example of the figure, the View Point Transformation would align the images of the teacher's arm and that of the student.

Another relevant aspect for the scope of this deliverable is the segmentation of the hand/arm of a demonstrator from images, as shown in the figure below.



The process proposed is based on a colour segmentation scheme which is trained to distinguish skin-colour from other objects. Morphological filtering is applied to select large skin-coloured blobs. After normalization with respect to size and orientation, the extracted hand images are used in a classification methodology to recognize hand gestures. Even though encouraging results have been produced in real-time, additional work is needed to improve accuracy and robustness.

Details of the entire methodology and experimental results are described in ANNEX I of this deliverable.

3. Dense depth maps from binocular disparity channels

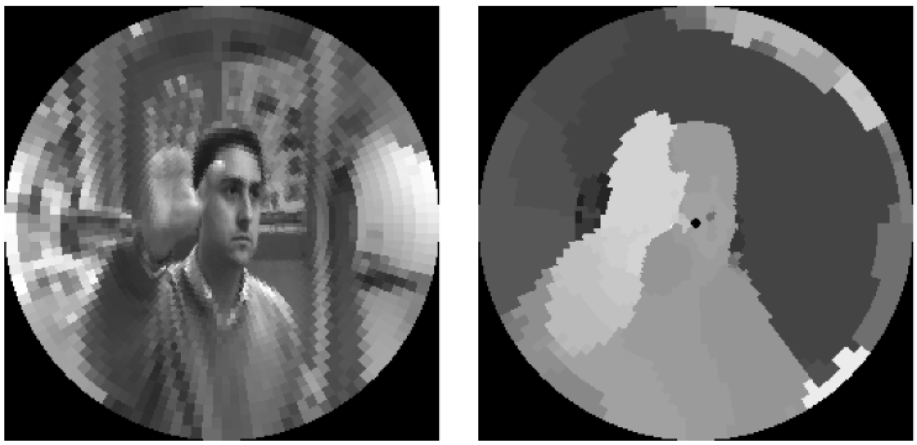
Foveation and stereopsis are important features on active vision systems. The former provides a wide field of view and high foveal resolution with low amounts of data, while the latter contributes to the acquisition of close range depth cues. The log-polar sampling has been proposed as an approximation to the foveated representation of the primate visual system. Although the huge amount of stereo algorithms proposed in the literature for conventional imaging geometries, very few are shown to work with foveated images sampled according to the log-polar transformation.

We have developed a method to extract dense disparity maps in real-time from a pair of log-mapped images, with direct application to active vision systems both for tracking or figure ground segmentation.

The method we propose uses the gray/color values of each pixel directly, without requiring any further feature extraction, making this method particularly suited for non-cartesian geometries, where the scale of analysis depends greatly on the variable to estimate (disparity). The fact that we obtain dense information has makes it suitable for object segmentation and region of interest selection.

The method starts by assuming a set of disparity hypotheses (mainly horizontal) and a probabilistic formulation is settled to determine the probability of observing a given disparity value at a given pixel subject to a disparity prior. The multiple hypotheses are then combined and the MAP estimate of the disparity computed.

There are inhibition links between different disparity channels for the same pixel (i.e. one pixel can only have one disparity value) and reinforcement between similar disparity values associated to neighbouring pixels. Since the MAP is only computed after these processes, the resulting disparity field preserves depth discontinuities, thus being very appropriate for figure-ground segmentation.



The picture above shows an example of one of the input images and the estimated disparity map. A detailed description and experimental results are included in ANNEX II.

4. Low-level image primitives

The following Visual Primitives acting at the lower level of the processing chain of the final system have been developed:

Log-Polar Mapping ActiveX Object (v2.0, 2002-10-25) - Allow to map Cartesian images to log-polar geometry. Various parameters of the log-polar map can be set using this software.

Corner Detection ActiveX Object (v1.1, 2002-01-08) - Extracts corners from images. The process is based on the Harris corner detector. It starts by computing the matrix of second derivatives of the image (Hessian) and the lowest eigenvalue is selected to indicate the lowest spatial curvature of the image patch. This value must be above a threshold for the feature extractor to select a certain point.

Corner Tracker ActiveX Object (v1.1, 2002-01-08) – Using the previous corner detector, this module tracks a corner over an image sequence by applying thewell-know Lucas-Kanade tracking algorithm. The algorithm is based on a first order approximation of an image pair and the optimal disparity is computed in a least-squares sense.

Normal Flow ActiveX Object (v1.0, 2002-03-28) – The normal flow corresponds to the projection of image motion on the direction of the image spatial gradient. It consists of the unique component of the image motion that can be computed locally, due to the well-known aperture problem. The normal flow is computed from spatio-temporal image derivatives.

Log-Polar Warping ActiveX Object (v2.0, 2002-10-25) - This software is an implementation of the process described in Section 3 of this deliverable. A set of disparity hypotheses are used to generate warped versions of log-polar images that are used in a probabilistic framework to obtain a dense disparity map of the observed scene.

5. ANNEXES

I - Visual Transformations in Gesture Imitation
II - A Binocular Stereo Algorithm for Log-polar Foveated Systems

Visual Transformations in Gesture Imitation: what you see is what you do

Manuel Cabido Lopes José Santos-Victor

Instituto de Sistemas e Robótica
Instituto Superior Técnico
Lisbon, Portugal
{macl,jasv}@isr.ist.utl.pt

Abstract

We propose an approach for a robot to imitate the gestures of a human demonstrator. Our framework consists solely of two components: a *Sensory-Motor Map* (SMM) and a *View-Point Transformation* (VPT). The SMM establishes an association between an arm image and the corresponding joint angles and it is learned by the system during a period of observation of its own gestures. The VPT is widely discussed in the psychology of visual perception and is used to transform the image of the demonstrator's arm to the so-called ego-centric image, as if the robot were observing its own arm. Different structures of the SMM and VPT are proposed in accordance with observations in human imitation. The whole system relies on monocular visual information and leads to a parsimonious architecture for learning by imitation. Real-time results are presented and discussed.

1 Introduction

The impressive advance of research and development in robotics and autonomous systems in the past years has led to the development of robotic systems of increasing motor, perceptual and cognitive capabilities.

These achievements are opening the way for new application opportunities that will require these systems to interact with other robots or non technical users during extended periods of time. Traditional programming methodologies and robot interfaces will no longer suffice, as the system needs to learn to execute complex tasks and improve its performance through its lifetime.

Our work has the long-term goal of building sophisticated robotic systems able to interact with humans or other robots in a natural and intuitive way. One promising approach relies on imitation whereby a robot could learn how to handle a person's private objects by observing the owner's behavior, over time.

Learning by imitation is not a new topic and has been addressed before in the literature. This learning paradigm has already been pursued in humanoid robotic applications [1] where the number of degrees of freedom is very large, tele-operation [2] or assembly tasks [3]. Most published works, however, describe complete imitation systems but focus their attention on isolated system components only, while we describe a complete architecture.

We will concentrate on the simplest form of imitation that consists in replicating the gestures or movements of a demonstrator, without seeking to understand the gestures or the action's goal. In the work described in [4], the imitator can not only replicate the gestures but also the dynamics of a demonstrator, but it requires the usage of an exoskeleton to sense the demonstrator's behavior. Instead, our approach is exclusively based on vision.

The motivation to use visual information for imitation arises from the fact that many living beings - like humans - resort to vision to solve an extremely large set of tasks. Also, from the engineering point view, video cameras are low-cost, non invasive devices that can be installed in ordinary houses and that provide an enormous quantity of information, specially if combined with domain knowledge or stereo data.

Interestingly, the process of imitation seems to be the primary learning process used by infants and monkeys during the first years of life. Recently, the discovery of the *mirror neurons* in the monkey's brain [5, 6] has raised new hypotheses and provided a better understanding of the process of imitation in nature. These neurons are activated both when a monkey performs a certain action and when it sees the same action being performed by a demonstrator or another monkey.

Even if the role of these neurons is not yet fully understood, a few important conclusions can nevertheless be drawn. Firstly, mirror neurons clearly illustrate the intimate relationship between perception and action. Secondly, these neurons exhibit the remarkable ability of "recognizing" certain gestures or actions when seen from very different perspectives (associating gestures performed by the demonstrator to the subject's own gestures).

One of the main contributions of this paper is related to this last observation, that is illustrated in Figure 1. We propose a method that allows the system to "rotate" the image of gestures done by a demonstrator (allo-image) to the corresponding image (ego-image) that would be obtained if those same gestures were actually performed by the system itself. We call this process the *View-Point Transformation* (VPT). Surprisingly, in spite of the importance given to the VPT in psychological studies [7], it has received very little attention from other researchers in the field of visual imitation.

Figure 1: Gestures can be seen from very distinct perspectives. The image shows one's own arm performing a gesture (ego-image) and that of the demonstrator performing a similar gesture (allo-image).

One of the few works that dealt explicitly with the VPT is [8]. However, instead of considering the complete arm posture, only the mapping of the end-effector position is done. The map between the allo and ego image is performed using epipolar geometry, based on a stereo camera pair.

Other studies addressed this problem in an implicit and superficially way. A mobile robot capable of learning the policy followed by another mobile vehicle is described in [9]. Since the system kinematics is very simple, the VPT corresponds to a transformation between the views of the two mobile robots. This is achieved in practice by delaying the imitator's perception until it reaches the same place as the demonstrator, without focusing the process of VPT. The work described in [10] has similar objectives to our own research and allows a robot to mimic the "dance" of an Avatar. However, it does not address the VPT at all, and a special invasive hardware was used to perform this transformation. Instead, we present a simple architecture for imitation which carefully addresses the fundamental process of *View-Point Transformation*.

The VPT allows the robot to map observed gestures to a canonical point-of-view. The final step consists in transforming these mapped features to motor commands, which is referred to as the *Sensory-Motor Map* (SMM). Our complete architecture for imitation is shown in Fig. 2.

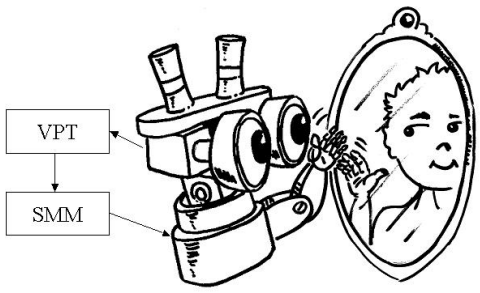


Figure 2: The combination of the Sensory-Motor Map and the View-Point Transformation allow the robot to imitate the arm movements executed by another robot or human.

The Sensory-Motor Map can be computed explicitly if the parameters of the arm-hand-eye configuration are known a priori but - more interestingly - it can be learned from observations of arm/hand motions. Again, biology can provide relevant insight. The Asymmetric Tonic Neck reflex [11] forces newborns to look to their hands, which allows them to learn the relationship between motor actions and the corresponding visual stimuli.

Similarly, in our work the robot learns the SMM during an initial period of self-observation, while performing hand/arm movements. Once the SMM has been estimated, the robot can observe a demonstrator, use the VPT to transform the image features to a canonical reference frame and map these features to motor commands through the SMM. The final result will be a posture similar to that observed.

Structure of the paper

In Section 2, we present the models used throughout this work, namely the arm kinematics and the camera/eye geometry. Section 3 is devoted to the definition and estimation of the *Sensory-Motor Map*. In Section 4 we describe how the system performs the *View-Point Transformation*. In Section 5 we show how to use these elementary blocks to perform imitation and present experimental results. In Section 6, we draw some conclusions and establish directions for future work.

2 Modeling

Throughout the paper we consider a robotic system consisting of a computer simulating an antropomorphic arm and equipped with a real web camera. This section presents the models used for the camera and robot body.

2.1 Body/arm kinematics

The anthropomorphic arm is modeled as an articulated link system. Fig. 3 shows the four arm links: L_1 - forearm, L_2 - upper arm, L_3 - shoulder width and L_4 - body height.

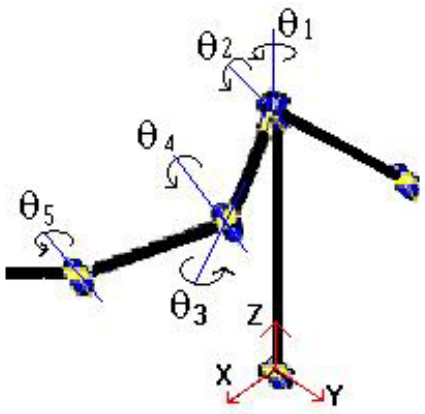


Figure 3: Kinematic model of the human arm.

It is further assumed that the relative sizes of these links are known, e.g. from biometric measurements: $L_1 = L_2 = 1$, $L_3 = 1.25$ and $L_4 = 2.5$.

2.2 Camera/eye geometry

An image is a projection of the $3D$ world whereby depth information is lost. In our case, we will retrieve depth information from a single image by using knowledge about the body links and a simplified, orthographic camera model.

We use the scaled orthographic projection model that assumes that the image is obtained by projecting all points along parallel lines plus a scale factor. Interestingly, such approximation may have some biological grounding taking into account the scale-compensation effect in the human vision [12] whereby we normalize the sizes of known objects irrespective to their distances to the eye.

Let $\mathbf{M} = [X\ Y\ Z]^T$ denote a 3D point expressed in the camera coordinate frame. Then, with an orthographic camera model, $\mathbf{M}$ is projected onto $\mathbf{m} = [u\ v]^T$, according to:

$$\begin{aligned} \mathbf{m} &= \mathcal{P}\mathbf{M} \\ \begin{bmatrix} u \\ v \end{bmatrix} &= s \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \end{aligned} \quad (1)$$

where s is a scale factor that can be estimated placing a segment with size L fronto-parallel to the camera and measuring the image size $l (s = l/L)$.

For simplification, we assume that the camera axis is positioned in the imitator's right shoulder with the optical axis pointing forward horizontally. With this specification of the camera pose, there is no need for an additional arm-eye coordinate transformation in Equation (1).

3 Sensory-Motor Map

The *Sensory-Motor Map* (SMM) defines a correspondence between perception and action. It can be interpreted in terms of forward/inverse kinematics for the case of robotic manipulators. The SMM can be used to predict the image resulting from moving one's arm to a certain posture. In our case, the SMM will allow the system to determine the arm's joint angles that correspond to a given image configuration of the arm.

In the context of imitation, the SMM can be used with different levels of ambiguity/completeness. In some cases, one wants to replicate exactly someone else's gestures, considering all the joint angles. In some other cases, however, we may want to imitate the hand pose only, while the position of the elbow or the rest of the arm configuration is irrelevant. To encompass these possibilities, we have considered two cases: the *full arm SMM* and the *free-elbow SMM* that will be described in the following sections. Finally we describe how the system can learn the SMM during a period of self-observation.

3.1 Full-Arm SMM

We denote the elbow and wrist image coordinates by $\mathbf{m_e}$ and $\mathbf{m_w}$, the forearm and upper arm image length by l_1 and l_2 and $\theta_{i=1..4}$, the various joint angles. We then have:

$$[\theta_1, \cdots, \theta_4] = \mathcal{F}_1(\mathbf{m_e}, \mathbf{m_w}, l_1, l_2, L_1, L_2, s)$$

where $\mathcal{F}_1(.)$ denotes the SMM, L_2/L_1 represents the (known) length of the upper/forearm and s is the camera scale factor.

The computation of this function can be done in successive steps, where the angles of the shoulder joint are determined first and used in a later stage to simplify the calculation of the elbow joint's angles.

The inputs to the SMM consist of features extracted from the image points of the shoulder, elbow and wrist; the outputs are the angular positions of every joint. The shoulder pan and elevation angles, θ_1 and θ_2 can be readily obtained from image data as:

$$\begin{aligned} \theta_1 &= f_1(\mathbf{m_e}) = \arctan(v_e/u_e) \\ \theta_2 &= f_2(l_2, L_2, s) = \arccos(l_2/sL_2) \end{aligned}$$

Once the system has extracted the shoulder angles, the process is repeated for the elbow. Before computing this second set of joint angles, the image features undergo a set of transformations so as to

compensate the rotation of the shoulder:

$$\begin{bmatrix} u'_w \\ v'_w \\ \xi \end{bmatrix} = \mathcal{R}_{zy}(\theta_1, \theta_2) \left(\begin{bmatrix} u_w \\ v_w \\ \sqrt{s^2 L_1^2 - l_1^2} \end{bmatrix} - \begin{bmatrix} u_e \\ v_e \\ 0 \end{bmatrix} \right) \quad (2)$$

where ξ is not used in the remaining computations and $\mathcal{R}_{zy}(\theta_1, \theta_2)$ denotes a rotation of θ_1 around the z axis followed by a rotation of θ_2 around the y axis.

With the transformed coordinates of the wrist we can finally extract the remaining joint angles, θ_3 and θ_4:

$$\begin{aligned} \theta_3 &= f_3(\mathbf{m}'_\mathbf{w}) = \arctan(v'_w/u'_w) \\ \theta_4 &= f_4(\mathbf{m}'_\mathbf{w}, L1, s) = \arccos(l'_1/sL_1) \end{aligned}$$

The approach just described allows the system to determine the joint angles corresponding to a certain image configuration of the arm. In the next section, we will address the case where the elbow joint is allowed to vary freely.

3.2 Free-Elbow SMM

The *free-elbow* SMM is used to generate a given hand position, while the elbow is left free to reach different configurations. The input features consist of the hand image coordinates and the depth between the shoulder and the hand.

$$[\theta_1, \theta_2, \theta_4] = \mathcal{F}_2(\mathbf{m}_\mathbf{w}, {}^r dZ_w, L_1, L_2, s)$$

The elbow joint, θ_3, is set to a comfortable position. This is done in an iterative process aiming at maintaining the joint positions as far as possible from their limit values. The optimal elbow angle position, $\hat{\theta}_3$ is chosen to maximize:

$$\hat{\theta}_3 = \arg\max_{\theta_3} \sum_i (\theta_i - \theta_i^{limits})^2$$

while the other angles can be calculated from the arm features. Again, the estimation process can be done sequentially, each joint being used to estimate the next one:

$$\begin{aligned} \theta_4 &= \arcsin\left(\frac{{}^r x_h^2 + {}^r y_h^2 + {}^r z_h^2}{2} - 1 \right) \\ \theta_1 &= 2\arctan\left(\frac{b_1 - \sqrt{b_1^2 + a_1^2 - c_1^2}}{a_1 + c_1} \right) \\ \theta_2 &= 2\arctan\left(\frac{b_2 - \sqrt{b_2^2 + a_2^2 - c_2^2}}{a_2 + c_2} \right) + \pi \end{aligned}$$

where the following constants have been used:

$$\begin{aligned} a_1 &= \sin\theta_4 + 1 \\ b_1 &= \cos\theta_3 \cos\theta_4 \\ c_1 &= -{}^r y_h \\ a_2 &= \cos\theta_4 \cos\theta_2 \cos\theta_3 - \sin\theta_2(1 + \sin\theta_4) \\ b_2 &= -\cos\theta_4 \sin\theta_3 \\ c_2 &= {}^r x_h \end{aligned}$$

3.3 Learning the SMM

In the previous sections we have derived the expressions of the full-arm and free-elbow SMMs. However, rather than coding these expressions directly we adopted a learning approach whereby the system learns the SMM by performing arm movements and observing the effect on the image plane.

The computation of the SMM can be done sequentially: estimating the first angle, which is then used in the computation of the following angle and so forth. This fact allows the system to learn the SMM as a sequence of smaller learning problems.

This approach has strong resemblance to the development of sensory-motor coordination in newborns and young infants, which starts by simple motions that get more and more elaborate as infants acquire a better control over motor coordination.

In all cases, we use a *Multi-Layer Perceptron* (MLP) to learn the SMM, i.e. to approximate functions $f_{i,i=1..4}$. Table 1 presents the learning error and illustrates the good performance of our approach for estimating the SMM.

θ_1	θ_2	θ_3	θ_4
$3.6e^{-2}$	$3.6e^{-2}$	3.6	3.6

Table 1: Mean squared error (in deg.^{2}) for the each joint in the *full-arm SMM*

Ideas about development can be further exploited in this construction. Starting from simpler cases, de-coupling several degrees of freedom, interleaving perception with action learning cycles are developmental "techniques" found in biological systems.

4 View-Point Transformation

A certain arm gesture can be seen from very different perspectives depending on whether the gesture is performed by the robot (self-observation) or by the demonstrator.

One can thus consider two distinct images: the *ego-centric* image, I_a, during self-observation and the *allo-centric* image, I_e, when looking at other robots/people. The *View-Point Transformation* (VPT) has the role of aligning the allo-centric image of the demonstrator's arm, with the ego-centric image, as if the system were observing its own arm.

The precise structure of the VPT is related to the ultimate meaning of imitation. Experiments in psychology show that imitation tasks can be ambiguous. In some cases, humans imitate only partially the gestures of a demonstrator (e.g. replicating the hand pose but having a different arm configuration, as in sign language), use a different arm or execute gestures with distinct absolute orientations [13]. In

some other cases, the goal consists in mimicking someone else's gestures as completely as possible, as when performing dancing or dismounting a complex mechanical part.

According to the structure of the chosen VPT, a class of imitation behaviors can be generated. We consider two different cases. In the first case - 3D VPT - a complete three-dimensional imitation is intended. In the second case - 2D VPT - the goal consists in achieving coherence only in the image, even if the arm pose might be different. Depending on the desired level of coherence (2D/3D) the corresponding (2D/3D) VPT allows the robot to transform the image of an observed gesture to an equivalent image as if the gesture were executed by the robot itself.

4.1 3D View-Point Transformation

In this approach we explicitly reconstruct the posture of the observed arm in 3D and use fixed points (shoulders and hip) to determine the rigid transformation that aligns the allo-centric and ego-centric image features: We then have:

$$I_e = \mathcal{P}\, T\, Rec(I_a) = VPT(I_a)$$

where T is a 3D rigid transformation and $Rec(I_a)$ stands for the 3D reconstruction of the arm posture from allo-centric image features. Posture reconstruction and the computation of T are presented in the following sections.

4.1.1 Posture reconstruction

To reconstruct the 3D posture of the observed arm, we will follow the approach suggested in [14], based on the orthographic camera and articulated arm models presented in Section 2.

Let M_1 and M_2 be the 3D endpoints of an arm-link whose image projections are denoted by $\mathbf{m_1}$ and $\mathbf{m_2}$. Under orthography, the X, Y coordinates are readily computed from image coordinates (simple scale). The depth variation, $dZ = Z_1 - Z_2$, can be determined as:

$$dZ = \pm\sqrt{L^2 - \frac{l^2}{s^2}}$$

where $L = \|M_1 - M_2\|$ and $l = \|m_1 - m_2\|$.

If the camera scale factor s is not known beforehand, one can use a different value provided that the following constraint, involving the relative sizes of the arm links, is met:

$$s \geq \max_i \frac{l_i}{L_i} \quad i = 1..4 \tag{3}$$

Fig. 4 illustrates results of the reconstruction procedure. It shows an image of an arm gesture and the corresponding 3D reconstruction, achieved with a single view and considering that s and the arm links proportions were known.

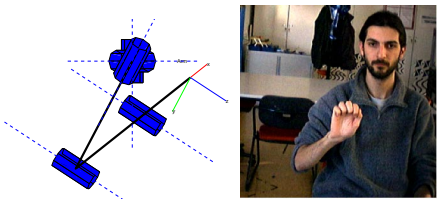


Figure 4: Left: Reconstructed arm posture. Right: Original view.

With this method there is an ambiguity in the sign of dZ. We overcome this problem by restricting the working volume of the arm. In the future, we will further address this problem and several approaches may be used: (i) optimization techniques to fit the arm kinematic model to the image; (ii) explore occlusions to determine which link is in the foreground; or (iii) use kinematics constraints to prune possible arm configurations.

4.1.2 Rigid Transformation (T)

A 3D rigid transformation is defined by three angles for the rotation and a translation vector. Since the arm joints are moving, they cannot be used as reference points. Instead, we consider the three points in Fig. 3: left and right shoulders, (M_{ls}, M_{rs}) and hip, M_{hip}, with image projections denoted by $(m_{ls}, m_{rs}, m_{hip})$. The transformation T is determined to translate and rotate these points until they coincide with those of the system's own body.

The translational component must place the demonstrators right shoulder at the image origin (which coincide's with the system's right shoulder) and can be defined directly in image coordinates:

$$t = -{}^a m_{rs}$$

After translating the image features directly, the remaining steps consist in determining the rotation angles to align the shoulder line and the shoulder-hip contour. The angles of rotation along the z, y and x axes, denoted by ϕ, θ and ψ are given by:

$$\begin{aligned} \phi &= \arctan\left(v_{ls}/u_{ls}\right) \\ \theta &= \arccos\left(u_{hip}/L_4\right) \\ \psi &= \arccos\left(v_{hip}/L_3\right) \end{aligned}$$

Hence, by performing the image translation first and the 3D rotation described in this section, we complete the process of aligning the image projections of the shoulders and hip to the ego-centric image coordinates.

4.2 $2D$ View-Point Transformation

The $2D$ VPT is used when one is not interested in imitating the depth variations of a certain move-

ment, alleviating the need for a full 3D transformation. It can also be seen as a simplification of the 3D VPT if one assumes that the observed arm describes a fronto-parallel movement with respect to the camera.

The $2D$ VPT performs an image translation to align the shoulder of the demonstrator (${}^{a}m_s$) and that of the system (at the image origin, by definition). The VPT can be written as:

$$VPT({}^{a}m) = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} [{}^{a}m - {}^{a}m_s] \qquad (4)$$

and is applied to the image projection of the demonstrator's hand or elbow, ${}^{a}m_h$ or ${}^{a}m_e$.

Notice that when the arm used to imitate is the same as the demonstrator, the imitated movement is a mirror image of the original. If we use a simple identity matrix in Equation (4) then the movement will be correct. At the image level both the 2D and 3D VPTs have the same result but the 3D posture of the arm is different in the two cases.

From the biological standpoint, the 2D VPT is more plausible than the 3D version. In [13] several imitation behaviors are presented which are not always faithful to the demonstrated gesture: sometimes, people do not care about usage of the correct hand, depth is irrelevant in some other cases, movements can be reflections of the original ones, etc. The 3D VPT might be more useful in industrial facilities where gestures should be reproduced as exactly as possible.

5 Experiments

We have implemented the modules discussed in the previous sections to build a system able to learn by imitation. In all the experiments, we use a web camera to observe the demonstrator gestures and a simulated robot arm to replicate those gestures.

We start by describing the approach used for hand-tracking before presenting the overall results of imitation. The position of the shoulder is assumed to be fixed. In the following sections we shall discuss about the procedures for doing imitation.

5.1 Hand Color Segmentation

To find the hand in the image we use a color segmentation scheme, implemented by a feed-forward neural network with three neurons in the hidden layer. As inputs we use the hue and saturation channels of HSV color representation. The training data are obtained by selecting the hand and the background in a sample image. After color classification a *majority* morphological operator is used. The hand is identified as the largest blob found and its position is estimated over time with a Kalman filter. Figure 5 shows a typical result of this approach.

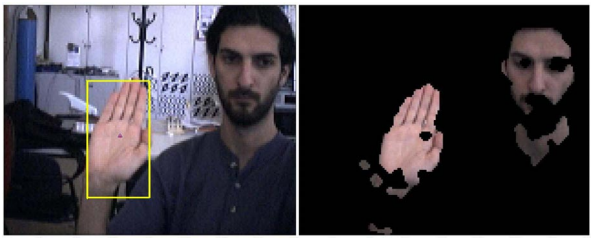


Figure 5: Skin color segmentation results.

5.2 Gesture Imitation

The first step to achieve imitation consists in training the system to learn the Sensory-Motor Map as described in Section 3.3. This is accomplished by a neural network that estimates the SMM while the system performs a large number of arm movements.

The imitation process consists of the following steps: (i) the system observes the demonstrator's arm movements; (ii) the VPT is used to transform these image coordinates to the *ego-image*, as proposed in Section 4 and (iii) the SMM generates the adequate joint angle references to execute the same arm movements.

Figure 6 shows experimental results obtained with the $3D$-VPT with the learned SMM (full-arm). To assess the quality of the results, we overlaid the images of the executed arm gestures (wire frame) on those of the demonstrator. The figure shows that the quality of imitation is very good.

Figure 7 shows results obtained in real-time (about 5 Hz) when using the $2D$ VPT and the *free-elbow* SMM. The goal of imitating the hand gesture is well achieved but, as expected, there are differences in the configuration of the elbow, particularly at more extreme positions.

These tests show that encouraging results can be obtained with the proposed framework under realistic conditions.

6 Conclusions and future work

We have proposed an approach for learning by imitation that relies exclusively on visual information provided by a single camera.

One of the main contributions is the *View-Point Transformation* that performs a "mental rotation" of the image of the demonstrator's arm to the *ego-image*, as if the system were observing its own arm. In spite of the fundamental importance of the VPT in visual perception and in the psychology of im-

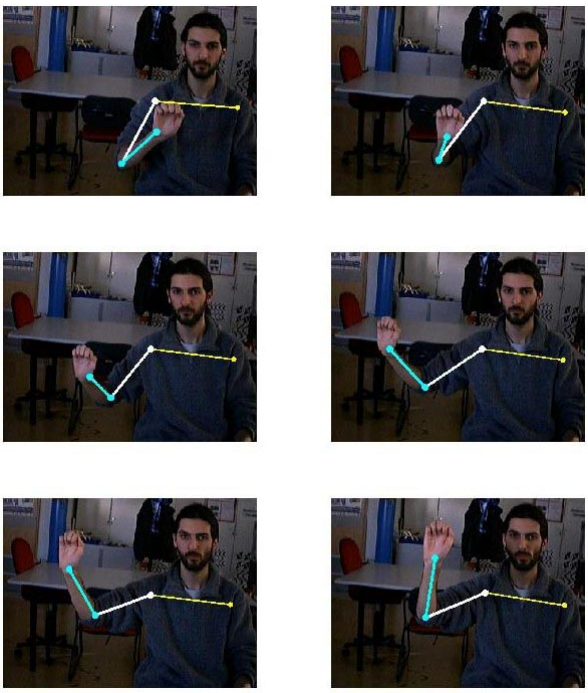


Figure 6: The quality of the results can be assessed by the coincidence of the demonstrator gestures and the result of imitation.

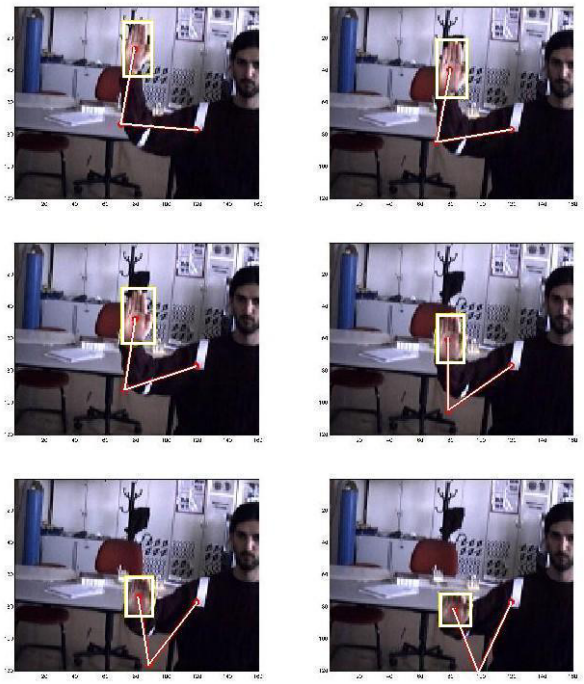


Figure 7: Family of solutions with different elbow angles, while the hand is faithfully imitated.

itation [7], it has received little attention by researchers in robotics.

We described two different VPTs needed for 3D or 2D imitation. The *View-Point Transformation* can have an additional interest to *Mirror Neurons* studies, by providing a canonical frame of reference that greatly simplifies the recognition of arm gestures.

The observed actions are mapped into muscles torques by the *Sensory-Motor Map*, that associates image features to motor acts. Again two different types of SMM are proposed, depending on whether the task consists of imitating the entire arm or the hand position only. The SMM is learned automatically during a period of self-observation.

Experiments conducted to test the various subsystems have led to encouraging results, thus validating our approach to the problem.

Besides improvements on the feature detection component using shape and kinematic data, future work will focus on the the understanding of the task goals to enhance the quality of imitation.

References

[1] S. Schaal. Is imitation learning the route to humanoid robots. *Trends in Cognitive Sciences*, 3(6):233–242, 1999.

[2] J. Yang, Y. Xu, and C.S. Chen. Hidden markov model approach to skill learning and its application to telerobotics. *IEEE Transations on Robotics and Automation*, 10(5):621–631, October 1994.

[3] T. G. Williams, J. J. Rowland, and M. H. Lee. Teaching from examples in assembly and manipulation of snack food ingredients by robot. In *2001 IEEE/RSJ, International Conference on Intelligent Robots and Systems*, pages 2300–2305, Oct.29-Nov.03 2001.

[4] Aaron D'Souza, Sethu Vijayakumar, and Stephan Schaal. Learning inverse kinematics. In *International Conference on Intelligent Robots and Systems*, pages 298–303, Maui, Hawaii, USA, 2001.

[5] V.S. Ramachandran. Mirror neurons and imitation learning as the driving force behind the great leap forward in human evolution. *Edge*, 69, June 2000.

[6] Luciano Fadiga, Leonardo Fogassi, Vittorio Gallese, and Giacomo Rizzolatti. Visuomotor neurons: ambiguity of the discharge or 'motor' perception? *International Journal of Psychophysiology*, 35:165–177, 2000.

[7] J.S. Bruner. Nature and use of immaturity. *American Psychologist*, 27:687–708, 1972.

[8] Minoru Asada, Yuichiro Yoshikawa, and Koh Hosoda. Learning by observation without three-dimensional reconstruction. In *Intelligent Autonomous Systems (IAS-6)*, pages 555–560, 2000.

[9] A. Billard and G. Hayes. Drama, a connectionist architecture for control and learning in autonomous robots. *Adaptive Behaviour*, 7(1):35–63, 1999.

[10] Maja J. Matarić. Sensory-motor primitives as a basis for imitation: Linking perception to action and biology to robotics. In C. Nehaniv K. Dautenhahn, editor, *Imitation in Animals and Artifacts*. MIT Press, 2000.

[11] G. Metta, G. Sandini, L. Natale, and F.Panerai. Sensorimotor interaction in a developing robot. In *First International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, pages 18–19, Lund, Sweden, September 2001.

[12] Richard L. Gregory. *Eye and Brain, The Psychology of Seeing*. Princeton University Press, Princeton, New Jersey, 1990.

[13] Philippe Rochat. Ego function of early imitation. In Andrew N. Meltzoff and Wolfgang Prinz, editors, *The Imitative Mind*. Cambridge University Press, 2002.

[14] C. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *Computer Vision and Image Understanding*, 80:349–363, 2000.

A Binocular Stereo Algorithm for Log-polar Foveated Systems

Alexandre Bernardino and José Santos-Victor

Instituto Superior Técnico
ISR - Torre Norte, Piso 7
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
{alex,jasv}@isr.ist.utl.pt

Abstract. Foveation and stereopsis are important features on active vision systems. The former provides a wide field of view and high foveal resolution with low amounts of data, while the latter contributes to the acquisition of close range depth cues. The log-polar sampling has been proposed as an approximation to the foveated representation of the primate visual system. Although the huge amount of stereo algorithms proposed in the literature for conventional imaging geometries, very few are shown to work with foveated images sampled according to the log-polar transformation. In this paper we present a method to extract dense disparity maps in real-time from a pair of log-mapped images, with direct application to active vision systems.

1 Introduction

Stereoscopic vision is a fundamental perceptual capability both in animals and artificial systems. At close ranges, it allows reliable extraction of depth information, thus being suited for robotics tasks such as manipulation and navigation. In the last decades a great amount of research has been directed to the problem of extracting depth information from stereo imagery (see [25] for a recent review). However, the best performing techniques are still too slow to use on robotic systems which demand real-time operation. The straightforward way to reduce computation time is to work with coarse resolution images but this restricts the acquisition of detailed information all over the visual field. A better solution, inspired in biological systems, is the use of ocular movements together with foveated retinas. The visual system of primates has a space-variant nature where the resolution is high on the fovea (the center of the retina) and decreases gradually to the periphery of the visual field. This distribution of resolution is the evolutionary solution to reduce the amount of information traversing the optical nerve while maintaining high resolution in the fovea and a wide visual field. Moving the high resolution fovea we are able to acquire detailed representations of the surrounding environment. The excellent performance of biological visual systems led researchers to investigate the properties of foveated systems. Many active vision systems have adopted this strategy and since foveated images contain less information than conventional uniform resolution images, one obtains important reductions on the computation time.

2nd Workshop on Biologically Motivated Computer Vision, Tuebingen, Germany, Nov. 2002.

We may distinguish between two main methods to emulate foveated systems, that we denote by *multi-scale uniform sampling* methods and *non-uniform sampling* methods. *Uniform* methods preserve the cartesian geometry of the representation by performing operations at different scales in multi-resolution pyramids (e.g. [17],[10],[13]). Sampling grids are uniform at each level but different levels have different spacing and receptive field size. Notwithstanding, image processing operations are still performed on piecewise uniform resolution domains. *Non-uniform* methods resample the image with non-linear transformations, where receptive field spacing and size are non-uniform along the image domain. The VR transform [2], the DIEM method [19], and several versions of the *logmap* [30], are examples of this kind of methods.

The choice of method is a matter of preference, application dependent requirements and computational resources. *Uniform methods* can be easier to work with, because many current computer vision algorithms can be directly applied to these representations. However, *non-uniform* methods can achieve more compact image representations with consequent benefits in computation time. In particular the *logmap* has been shown to have many additional properties like rotation and scale invariance [31], easy computation of time-to-contact [28], improved linear flow estimation [29], looming detection [23], increased stereo resolution on verging systems [14], fast anisotropic diffusion [11], improved vergence control and tracking [7, 3, 4].

Few approaches have been proposed to compute disparity maps for foveated active vision systems, and existing ones rely on the foveated pyramid representation [17, 27, 6]. In this paper we describe a stereo algorithm to compute dense disparity maps on *logmap* based systems. Dense representations are advantageous for object segmentation and region of interest selection. Our method uses directly the gray/color values of each pixel, without requiring any feature extraction, making this method particularly suited for non-cartesian geometries, where the scale of analysis depends greatly on the variable to estimate (disparity).

To our knowledge, the only work to date addressing the computation of stereo disparity in *logmap* images is [15]. In that work, disparity maps are obtained by matching laplacian features in the two views (zero crossing), which results in sparse disparity maps.

2 Real-Time Log-polar Mapping

The log-polar transformation, or *logmap*, $\mathbf{l}(\mathbf{x})$, is defined as a conformal mapping from the *cartesian* plane $\mathbf{x} = (x, y)$ to the *log-polar* plane $\mathbf{z} = (\xi, \eta)$:

$$\mathbf{l}(\mathbf{x}) = \begin{bmatrix} \xi \\ \eta \end{bmatrix} = \begin{bmatrix} \log(\sqrt{x^2 + y^2}) \\ \arctan \frac{y}{x} \end{bmatrix} \tag{1}$$

Since the *logmap* is a good approximation to the retino–cortical mapping in the human visual system [26, 12], the cartesian and log-polar coordinates are also called "retinal" and "cortical", respectively. In continuous coordinates, a

cortical image I^{cort} is obtained from the corresponding retinal image I by the warping:

$$I^{cort}(\mathbf{z}) = I(\mathbf{l}^{-1}(\mathbf{x}))$$

A number of ways have been proposed to discretize space variant maps [5]. We have been using the *logmap* for some years in real-time active vision applications [3, 4]. To allow real-time computation of *logmap* images we partition the retinal plane into *receptive fields*, whose size and position correspond to a uniform partition of the cortical plane into *super-pixels* (see Fig. 1). The value of a *super-pixel* is given by the average of all pixels in the corresponding receptive field.

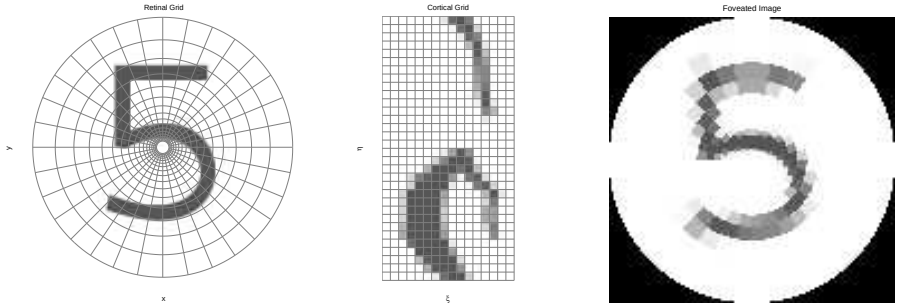


Fig. 1. The log-polar sampling scheme is implemented by averaging the pixels contained within each of the *receptive fields* shown in the left image. These space-variant receptive fields are angular sections of circular rings corresponding to uniform rectangular *super-pixels* in the cortical image (center). To reconstruct the retinal image, each receptive field gets the value of the corresponding *super-pixel* (right).

3 Disparity map computation

We start describing an intensity based method to find the likelihood of stereo matches in usual cartesian coordinates, $\mathbf{x} = (x, y)$. Then we show how the method can be extended to cope with *logmap* images. Finally we describe the remaining steps to obtain the disparity maps.

Let I and I' be the left and right images, respectively. For depth analysis, we are interested in computing the horizontal disparity map, but since we consider a general head vergence configuration, vertical disparities must also be accounted for. Therefore, disparity is a two valued function defined as $\mathbf{d}(\mathbf{x}) = (d_x, d_y)$. Taking the left image as the reference, the disparity at point $\mathbf{x}$ is given by $\mathbf{d}(\mathbf{x}) = \mathbf{x}' - \mathbf{x}$, where $\mathbf{x}$ and $\mathbf{x}'$ are the locations of matching points in the left and right images. If a pixel at location $\mathbf{x}$ in the reference image is not visible in the right image, we say the pixel is occluded and disparity is undefined ($\mathbf{d}(\mathbf{x}) = \varnothing$).

3.1 Bayesian formulation

To obtain dense representations, we use an intensity based method similar to [32]. We formulate the problem in a discrete bayesian framework. Having a finite set of possible disparities, $D = \{\mathbf{d}_n\}, n = 1 \cdots N$, for each location $\mathbf{x}$ we define a set of hypothesis, $H = \{h_n(\mathbf{x})\}, n = 0 \cdots N$, where $h_0(\mathbf{x})$ represents the occlusion condition ($\mathbf{d}(\mathbf{x}) = \varnothing$), and the other h_n represent particular disparity values, $\mathbf{d}(\mathbf{x}) = \mathbf{d}_n$. Other working assumptions are the following:

1. Object appearance does not vary with view point (lambertian surfaces) and cameras have the same gain, bias and noise levels. This corresponds to the *Brightness Constancy Assumption* [16]. Considering the existence of additive noise, we get the following stereo correspondence model:

$$I(\mathbf{x}) = I'(\mathbf{x} + \mathbf{d}(\mathbf{x})) + \eta(\mathbf{x}) \tag{2}$$

2. Noise is modeled as being independent and identically distributed with a certain probability density function, f. In the unoccluded case, the probability of a certain gray value $I(\mathbf{x})$ is conditioned by the value of the true disparity $\mathbf{d}(\mathbf{x})$ and the value of I' at position $\mathbf{x} + \mathbf{d}(\mathbf{x})$:

$$Pr(I(\mathbf{x})|\mathbf{d}(\mathbf{x})) = f(I(\mathbf{x}) - I'(\mathbf{x} + \mathbf{d}(\mathbf{x})))$$

We assume zero-mean gaussian white noise, and have $f(t) = 1/\sqrt{2\pi\sigma^2}e^{-t^2/2\sigma^2}$ where σ^2 is the noise variance.

3. In the discrete case we define the *disparity likelihood* images as:

$$L_n(\mathbf{x}) = Pr(I(\mathbf{x})|h_n(\mathbf{x})) = f(I(\mathbf{x}) - I'_n(\mathbf{x})) \tag{3}$$

where $I'_n(\mathbf{x}) = I'(\mathbf{x} + \mathbf{d_n})$ are called *disparity warped images.*

4. The probability of a certain hypothesis given the image gray levels (posterior probability) is given by the Bayes' rule:

$$Pr(h_n|I) = \frac{Pr(I|h_n)Pr(h_n)}{\sum_{i=0}^{N} Pr(I|h_i)Pr(h_i)} \tag{4}$$

where we have dropped the argument $\mathbf{x}$ since all functions are computed at the same point.

5. If a pixel at location $\mathbf{x}$ is occluded in the right image, its gray level is unconstrained and can have any value in the set of M admissible gray values,

$$Pr(I|h_0(\mathbf{x})) = \frac{1}{M} \tag{5}$$

We define a prior probability of occlusion with a constant value for all sites:

$$Pr(h_0) = q \tag{6}$$

6. We do not favor any *a priori* particular value of disparity. A constant prior is considered and its value must satisfy $Pr(h_n) \cdot N + q = 1$, which results in:

$$Pr(h_n) = (1 - q)/N \tag{7}$$

7. Substituting the priors (5), (6), (7), and the likelihood (3) in (4), we get:

$$Pr(h_n|I) = \begin{cases} \frac{L_n(I)}{\sum_{i=1}^{N} L_i(I)+qN/(M-qM)} & \Leftarrow n \neq 0 \\ \frac{qN/(M-qM)}{\sum_{i=1}^{N} L_i(I)+qN/(M-qM)} & \Leftarrow n = 0 \end{cases} \tag{8}$$

The choice of the hypothesis that maximizes (8) leads us to the MAP (*maximum a posteriori*) estimate of disparity[1]. However, without any further assumptions, there may be many ambiguous solutions. It is known that in the general case, the stereo matching problem is under-constrained and ill-posed [25]. One way to overcome this fact is to assume that the scene is composed by piece-wise smooth surfaces and introduce spatial interactions between neighboring locations to favor smooth solutions. Later we will describe a cooperative spatial facilitation method to address this problem.

3.2 Cortical Likelihood Images

While in cartesian coordinates the *disparity warped images* can be obtained by shifting pixels by an amount independent of position, $\mathbf{x}' = \mathbf{x} + \mathbf{d}_n$, in cortical coordinates the disparity shifts are different for each pixel, as shown in Fig.2. Thus, for each cortical pixel and disparity value, we have to compute the corresponding pixel in the second image. Using the *logmap* definition (1), the cortical correspondences can be obtained by:

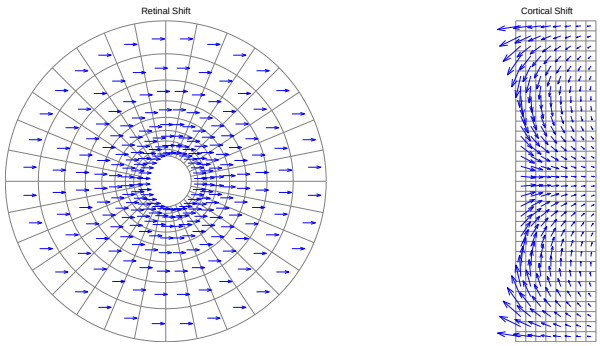


Fig. 2. A space invariant shift in retinal coordinates (left) corresponds to a space variant warping in the cortical array.

$$\mathbf{z}'_n(\mathbf{z}) = \mathbf{l}\Big(\mathbf{l}^{-1}(\mathbf{z}) + \mathbf{d_n}\Big) \tag{9}$$

This map can be computed off-line for all cortical locations and stored in a look-up table to speed-up on-line calculations. To minimize discretization errors, the

[1] The terms in the denominator are normalizing constants and do not need to be computed explicilty

weights for intensity interpolation can also be pre-computed and stored. A deeper explanation of this technique can be found in [22].

Using the pre-computed look up tables, the cortical *disparity warped images* can be efficiently computed on-line:

$$I_n^{cort'}(\mathbf{z}) = I^{cort'}(\mathbf{z}_n(\mathbf{z}))$$

From Eq. (3) we define $N+1$ *cortical likelihood images*, $L_n^{cort}(\mathbf{z})$, that express the likelihood of a particular hypothesis at cortical location $\mathbf{z}$:

$$L_n^{cort}(\mathbf{z}) = f(I^{cort}(\mathbf{z}) - I_n^{cort'}(\mathbf{z}))$$

Substituting this result in Eq. (8) we have the cortical posterior probabilities:

$$Pr^{cort}(h_n|I^{cort}) \propto \begin{cases} L_n^{cort}(I) & \Leftarrow n \neq 0 \\ qN/(M - qM) & \Leftarrow n = 0 \end{cases} \tag{10}$$

3.3 Cooperative spatial facilitation

The value of the likelihood images L_n^{cort} at each cortical location $\mathbf{z}$ can be interpreted as the response of disparity selective neurons, expressing the degree of match between corresponding locations in the right and left images. When many disparity hypothesis are likely to occur (e.g. textureless areas) several neurons tuned to different disparities may be simultaneously active. In a computational framework, this "aperture" problem is usually addressed by allowing neighborhood interactions between units, in order to spread information from and to non-ambiguous regions. A bayesian formulation of these interactions leads to Markov Random Fields techniques [33], whose existing solutions (annealing, graph optimization) are still computationally expensive. Neighborhood interactions are also very commonly found in biological literature and several cooperative schemes have been proposed, with different facilitation/inibhition strategies along the spatial and disparity coordinates [18, 21, 20]. For the sake of computational complexity we adopt a spatial-only facilitation scheme whose principle is to reinforce the output of units at locations whose coherent neighbors (tuned for the same disparity) are active. This scheme can be implemented very efficiently by convolving each of the *cortical likelihood images* with a low-pass type of filter, resulting on $N + 1$ *Facilitated Cortical Likelihood Images*, F_n^{cort}. We use a fast IIR isotropic separable first order filter, which only requires two multiplications and two additions per pixel. We prefer filters of large impulse response, which provide better smoothness properties and favor blob like objects, at the cost of missing small or thin structures in the image. Also, due to the space-variant nature of the cortical map, regions on the periphery of the visual field will have more "smoothing" than regions in the center.

At this point, it is worth noticing that since the 70's, biological studies show that neurons tuned to similar disparities are organized in clusters on visual cortex area V2 in primates [8], and more recently this organization has also been found on area MT [9]. Our architecture, composed by topographically organized maps of units tuned to the same disparity, agrees with these biological findings.

3.4 Computing the solution

Replacing in (10) the *cortical likelihood images* L_n^{cort} by their filtered versions F_n^{cort} we obtain $N+1$ *cortical disparity activation images*:

$$D_n^{cort} = \begin{cases} F_n^{cort}(I) & \Leftarrow n \neq 0 \\ qN/(M - qM) & \Leftarrow n = 0 \end{cases} \tag{11}$$

The disparity map is obtained by computing the hypothesis that maximizes the *cortical disparity activation* images for each location:

$$\hat{\mathbf{d}}(\mathbf{z}) = \arg\max_n (D_n^{cort}(\mathbf{z}))$$

In a neural networks perspective, this computation is analogous a winner-take-all competition between non-coherent units at the same spatial location, promoted by the existence of inhibitory connections between them [1].

4 Results

We have tested the proposed algorithm on a binocular active vision head in general vergence configurations, and on standard stereo test images. Results are shown on Figs. 3 and 4. Bright and dark regions correspond to near and far objects, respectively. The innermost and outermost rings present some noisy disparity values due to border effects than can be easily removed by simple post-processing operations.

Fig. 3. The image in the right shows the raw foveated disparity map computed from the pair of images shown in the left, taken from a stereo head verging on a point midway between the foreground and background objects.

Some intermediate results of the first experiment are presented in Fig. 5, showing the output of the cortical likelihood and the cortical activation for a particular disparity hypothesis. In the likelihood image notice the great amount of noisy points corresponding to false matches. The spatial facilitation scheme and the maximum computation over all disparities are essential to reject the false matches and avoid ambiguous solutions.

A point worth of notice is the blob like nature of the detected objects. As we have pointed out in section 3.3, this happens because of the isotropic nature

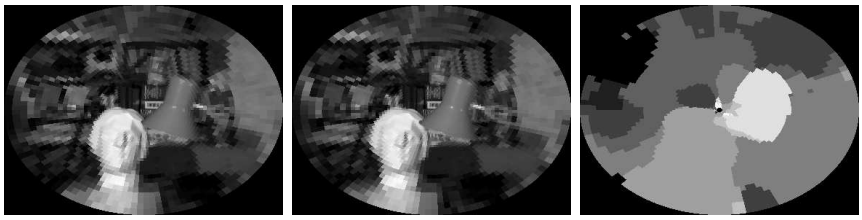


Fig. 4. The disparity map on the right was computed from the well known stereo test images from Tsukuba University. In the left we show the foveated images of the stereo pair. Notice that much of the detail in the periphery is lost due to the space variant sampling. Thus, this result can not be directly compared with others obtained from uniform resolution images.

and large support of the spatial facilitation filters. Also, the space variant image sampling, blurs image detail in the periphery of the visual field. This results in the loss of small and thin structures like the fingertips in the stereo head example and the lamp support in the Tsukuba images. However note that spatial facilitation do not blur depth discontinuities because filtering is not performed on the disparity map output, but on the likelihood maps before the maximum operation.

The lack of detail shown in the computed maps is not a major drawback for our applications, that include people tracking, obstacle avoidance and region of interest selection for further processing. As a matter of fact, it has been shown in a number of works that many robotics tasks can be performed with coarse sensory inputs if combined with fast control loops [24].

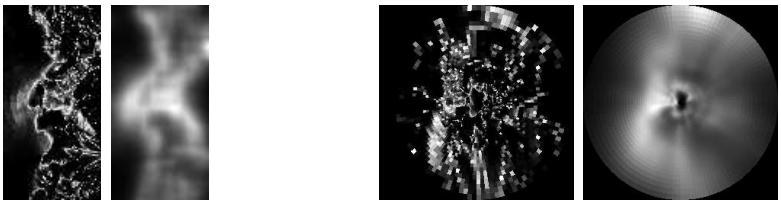


Fig. 5. Intermediate results for the experiment in Fig.3. This figure shows the cortical maps tuned to retinal disparity $d_i = 26$, for which there is a good match in the hand region. In the left group we show the likelihood images L_i^{cort} (left) and D_i^{cort} (right) corresponding to the cortical activation before and after the spatial facilitation step. In the right group, the same maps are represented in retinal coordinates, for better interpretation of results.

The parameters used in the tests are the following: log-polar mapping with 128 angular sections and 64 radial rings; retinal disparity range from -40 to 40 pixels (horizontal) and from -6 to 6 pixels (vertical), both in steps of 2; $q = 0.1$

(prior probability of occlusion); $M = 256$ (number of gray values); $\sigma = 3$ (white noise standard deviation); facilitation filtering with zero-phase forward/reverse filter $y(n) = 0.8y(n-1) + 0.2x(n)$.

The algorithms were implemented in C++ and take about three seconds to run in a PII 350MHz computer.

5 Conclusions

We have presented a real-time dense disparity estimation algorithm for foveated systems using the *logmap*. The algorithm uses an intensity based matching technique, which makes it easily extensible to other space variant sampling schemes. Some results were taken from an active stereo head and others obtained from standard test images. Many robots are currently equipped with foveated active vision systems and the availability of fast stereopsis will drastically improve their perceptual capabilities. Obstacle detection and tracking, region of interest selection and object manipulation are some possible applications.

Acknowledgements

This work was partially supported by EU project MIRROR: *Mirror Neurons based Robot Recognition, IST-2000-28159.*

References

1. S. Amari and M. Arbib. *Competition and Cooperation in Neural Nets*, pages 119–165. Systems Neuroscience. J. Metzler (ed), Academic Press, 1977.
2. A. Basu and K. Wiebe. Enhancing videoconferencing using spatially varying sensing. *IEEE Trans. on Systems, Man, and Cybernetics*, 38(2):137–148, Mar. 1998.
3. A. Bernardino and J. Santos-Victor. Binocular visual tracking : Integration of perception and control. *IEEE Trans. on Robotics and Automation*, 15(6):137–146, Dec. 1999.
4. A. Bernardino, J. Santos-Victor, and G. Sandini. Foveated active tracking with redundant 2d motion parameters. *Robotics and Autonomous Systems*, 39(3-4):205–221, June 2002.
5. M. Bolduc and M. Levine. A review of biologically motivated space-variant data reduction models for robotic vision. *CVIU*, 69(2):170–184, Feb. 1998.
6. T. Boyling and J. Siebert. A fast foveated stereo matcher. In *Proc. Conf. on Imaging Science Systems and Technology*, pages 417 – 423, Las Vegas, USA, 2000.
7. C. Capurro, F. Panerai, and G. Sandini. Dynamic vergence using log-polar images. *IJCV*, 24(1):79–94, Aug. 1997.
8. T. Wiesel D. Hubel. Stereoscopic vision in macaque monkey. cells sensitive to binocular depth in area 18 of the macaque mokey cortex. *Nature*, 225:41–42, 1970.
9. G. DeAngelis and W. Newsome. Organization of disparity–selective neurons in macaque area mt. *The Journal of Neuroscience*, 19(4):1398–1415, 1999.
10. S. Mallat E. Chang and C. Yap. Wavelet foveation. *J. Applied and Computational Harmonic Analysis*, 9(3):312–335, Oct. 2000.

11. B. Fischl, M. Cohen, and E. Schwartz. Rapid anisotropic diffusion using space-variant vision. *IJCV*, 28(3):199–212, July/Aug. 1998.
12. G. Gambardella G. Sandini, C. Braccini and V. Tagliasco. A model of the early stages of the human visual system: Functional and topological transformation performed in the peripheral visual field. *Biological Cybernetics*, 44:47–58, 1982.
13. W. Geisler and J. Perry. A real-time foveated multi-resolution system for low-bandwidth video communication. In *Human Vision and Electronic Imaging, SPIE Proceedings 3299*, pages 294–305, Aug. 1998.
14. N. Griswald, J. Lee, and C. Weiman. Binocular fusion revisited utilizing a log-polar tessellation. *CVIP*, pages 421–457, 1992.
15. E. Grosso and M. Tistarelli. Log-polar stereo for anthropomorphic robots. In *Proc. 6th ECCV*, pages 299 – 313, Dublin, Ireland, June-July 2000.
16. B. Horn. *Robot Vision*. MIT Press, McGraw Hill, 1986.
17. W. Klarquist and A. Bovik. Fovea: A foveated vergent active stereo system for dynamic three-dimensional scene recovery. *IEEE Trans. on Robotics and Automation*, 14(5):755 – 770, Oct. 1998.
18. D. Marr and T. Poggio. Cooperative computation of stereo disparity. *Science*, 194:283–287, 1976.
19. M. Peters and A. Sowmya. A real-time variable sampling technique: Diem. In *Proc. ICPR*, pages 316–321, Brisbane, Australia, Aug. 1998.
20. S. Pollard, J. Mayhew, and J. Frisby. Pmf: A stereo correspondence algorithm using a disparity gradient limit. *Perception*, 14:449–470, 1985.
21. K. Prazdny. Detection of binocular disparities. *Biol. Cybern*, 52:93–99, 1985.
22. G. Metta R. Manzotti, A. Gasteratos and G. Sandini. Disparity estimation on log-polar images and vergence control. *CVIU*, 83:97–117, 2001.
23. G. Salgian and D. Ballard. Visual routines for vehicle control. In D. Kriegman, G. Hager, and S. Morse, editors, *The Confluence of Vision and Control*. Springer Verlag, 1998.
24. J. Santos-Victor and A. Bernardino. Vision-based navigation, environmental representations, and imaging geometries. In *Proc. 10th Int. Symp. of Robotics Research*, Victoria, Australia, Nov. 2001.
25. D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1):7–42, April-June 2002.
26. E. Schwartz. Spatial mapping in the primate sensory projection : Analytic structure and relevance to perception. *Biological Cybernetics*, 25:181–194, 1977.
27. J. Siebert and D. Wilson. Foveated vergence and stereo. In *Proc. of the 3rd Int. Conf. on Visual Search (TICVS)*, Nottingham, UK, Aug. 1992.
28. M. Tistarelli and G. Sandini. On the advantages of polar and log-polar mapping for direct estimation of the time-to-impact from optical flow. *IEEE Trans. on PAMI*, 15(8):401–411, April 1993.
29. H. Tunley and D. Young. First order optic flow from log-polar sampled images. In *Proc. ECCV*, pages A:132–137, 1994.
30. R. Wallace, P. Ong, B. Bederson, and E. Schwartz. Space variant image processing. *IJCV*, 13(1):71–90, Sep. 1995.
31. C. Weiman and G. Chaikin. Logarithmic spiral grids for image processing and display. *Comp Graphics and Image Proc*, 11:197–226, 1979.
32. R. Zabih Y. Boykov, O. Veksler. Disparity component matching for visual correspondence. In *Proc. CVPR*, pages 470–475, 1997.
33. R. Zabih Y. Boykov, O. Veksler. Markov random fields with efficient approximations. In *Proc. CVPR*, pages 648–655, 1998.