Map Building from Human-Computer Interactions

The Author
Institution
First line of institution address
author@institution.com

Abstract

On-line help from a human actor will be exploited to facilitate computer perception. This paper proposes an innovative real-time algorithm – running on an active vision head – to build 3D scene descriptions from human cues. The theory is supported by experimental results both for figure/ground segregation of typical heavy objects in a scene (such as furniture), and for 3D object/scene reconstruction.

1. Introduction

Embodied and situated perception [4] consists of boosting the vision capabilities of an artificial creature by fully exploiting the opportunities created by an embodied agent situated in the world [2].

Active vision proponents [1, 5], contrary to passive vision, argue for active control of the visual perception mechanism so that perception is facilitated. Percepts can indeed be acquired in a purposive way by the active control of a camera [1]. This approach has been successfully applied to several computer vision problems, such as stereo vision - by dynamically changing the baseline distance between the cameras or by active focus selection [9].

We argue for solving a visual problem by not only actively controlling the perceptual mechanism, but also and foremost actively changing the environment through experimental manipulation. The human body plays an essential role in such a framework, being applied not only to facilitate perception, but also to change the world context so that percepts are easily understood [4].

1.1 Motivation

Besides binocular cues, the human visual system also processes monocular data for depth inference, such as focus, perspective distortion, among others. Previous attempts have been made on exploring scene context for depth inference [14]. However, these passive techniques make use of contextual clues already present on the scene. They do not actively change the context of the scene through manipulation to improve the robot's perception. We propose an active, embodied approach that actively changes the context of a scene, extracting monocular depth measures.

This paper proposes an algorithm to infer depth and build 3-dimensional maps from a distinct monocular cue: the relative size of objects on a monocular image – special focus will be placed on using the human's arm as a reference measure. Another algorithm's novelty is the real-time transmission of world-structure to the perceptual system from the action of an embodied agent (the human tutor). This real-time algorithm builds scene descriptions as a function of objects, together with 3D coarse maps for the scene, through the analysis of cues provided by an interacting human.

It should be emphasized we will not argue for more accurate results than other Stereo or Monocular depth inference techniques. By the contrary, the technique here proposed provides solely coarse depth information. Its power relies on providing an additional cue for depth inference, which could be augmented by using cues from other scene objects besides the human arm. In addition, the proposed algorithm has complementary properties to other depth inference algorithms, it does not require special hardware (low-cost cameras will suffice) and it also outputs object segmentations.

1.2 Human-Robot Interactive Communication

Previous approaches for transferring skills from human to computers rely heavily on human gesture recognitio, or haptic interfaces for detecting human motion. Environments are often over-simplified to facilitate the perception of the task sequence [10]. Other approaches consist of visually identifying simple guiding actions (such as direction following, or collision), for which both the task's structure and goal are well known [11].

Teaching a visual system information concerning the surrounding world is a difficult task, which takes several years for a child, equipped with evolutionary mechanisms stored in its genes, to accomplish. Our approach exploits help from a human in a robot's learning loop to extract meaningful percepts from the world. However, it should be emphasized that such help does not include constraining the world structure (for instance by removing environment cluttering or careful luminosity setup). The focus will be placed on communicating information to a robot which boosts its perceptual skills, helping the visual system to filter out irrelevant information. Indeed, while teaching a toddler, parents do not remove the room's furniture or buy extra lights to just show the child a book. Help instead is given by facilitating the child's task of stimulus selection (for example, by pointing or tapping into a book's image [4]).

1.3 Map Building

Several techniques have been proposed for three-dimensional reconstruction of environments, ranging from passive sensing techniques to active sensing using laser range finders, or both [13]. This paper will focus on learning topological map representations [6] from cues provided by interactive humans.

2. Object Segmentation from Human-Robot Interactive Cues

Real-time object segmentation on unstructured, non-static, noisy and low resolution (128×128) images is a hard problem, subject to a large variety of disturbances,

- ▷ target object with similar color/texture as background
- ▷ multiple objects moving simultaneously in a scene
- ▷ object is the union of a large number of color regions

Robustness to luminosity and world structure variations is also of paramount importance. Mobility constraints (such as segmenting heavy objects) poses additional difficulties, since motion cannot be used to facilitate the problem.

We argue for a visual embodied strategy which is not limited to active robotic heads. Instead, embodiment of an agent is exploited by probing the world with a human arm. This strategy proves not only useful to segment object descriptions from books [3], but also to segment large, stationary objects (such as a table) from monocular images.

2.1. Figure-Ground Segregation

We propose a human aided object segmentation algorithm to tackle the figure-ground problem. Indeed, a significant amount of contextual information may be extracted from a periodically moving actuator. This can be framed as the problem of estimating $p(o_n|v_{B_{\vec{p},\epsilon}}, act^{per}_{\vec{p},S})$, the probability of finding object n given a set of local, stationary features v on a neighborhood ball B of radius ϵ centered on location p, and a periodic actuator on such neighborhood with trajectory points in the set $S \subseteq B$. The following algorithm implements the estimation process to solve this figure-ground separation problem (see Figure 1):

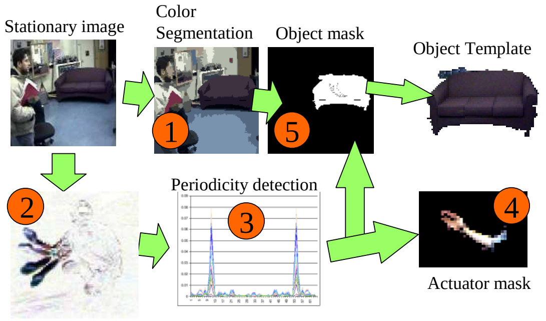


Figure 1. Segmentation of heavy, stationary objects. A standard color segmentation algorithm computes a compact cover of color clusters for the image. A human actor ***shows*** **the sofa to the robot, by waving on the objects' surface. The human actuator's periodic trajectory is used to extract the object's compact cover – the collection of color cluster sets which composes the object.**

1. A standard color segmentation [7] algorithm is applied to a stationary image
2. A human actor waves an arm on top of the target object
3. The motion of skin-tone pixels is tracked over a time interval (by the Lucas-Kanade Pyramidal algorithm). The energy per frequency content – using Short-Time Fourier Transform (STFT) – is determined for each point's trajectory
4. Periodic, skin-tone points are grouped together into the arm mask [4]
5. The trajectory of the arm's endpoint describes an algebraic variety [8] over N^2 (N stands for natural integers). The target object's template is then given by the union of all bounded subsets (the color regions of the stationary image) which intersect this variety

Periodic detection is applied at multiple scales. since the movement might not appear periodic at a coarser scale, but appear as such at a finer scale. If a strong periodicity is not found at a larger scale, the window size is halved and the procedure is repeated again. Periodicity is estimated from a periodogram built for all signals from the energy of the

STFTs over the frequency spectrum. These periodograms are processed by a collection of narrow bandwidth band-pass filters. Periodicity is found if, compared to the maximum filter output, all remaining outputs are negligible.

The algorithm consists of grouping together the colors that form an object. This grouping works by having periodic trajectory points being used as seed pixels. The algorithm fills the regions of the color segmented image whose pixel values are closer to the seed pixel values, using a 8-connectivity strategy. Therefore, points taken from waving are used to both select and group a set of segmented regions into the full object.

2.1.1 Results

Considering Figure 2, both sofa and table segmentations are hard cases to solve. The clustering of regions by table-like color content produces two disjoint regions. One of them corresponds to the table, but it is not possible to infer which just from the color content. But a human teacher can *show* the table to the robot by waving on the table's surface. The arm trajectory then links the table to the correct region. For the sofa case, segmentation is hard because the sofa appearance consists of a collection of color regions. It is necessary additional information to group such regions without including the background. Once more, a human tutor *describes* the object, so that the arm trajectory groups several color regions into the same object - the sofa.

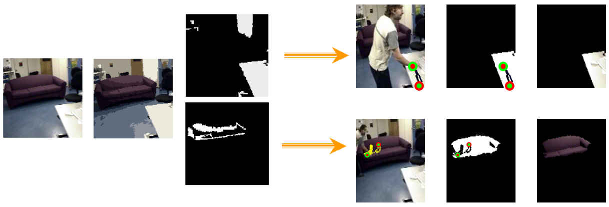


Figure 2. Segmentation of heavy, stationary objects. The arm trajectory links the objects to the correspondent color regions.

Figure 3 shows segmentations for a random sample of objects segmentations (furniture items), together with statistical results for such objects. Clusters grouped by a single trajectory might either form (eg. table) or not form (eg. black chair – a union of two disconnected regions) the smallest compact cover which contains the object (depending on intersecting or not all the clusters that form the object). After the detection of two or more temporally and spatially closed trajectories this problem vanishes – the black chair is grouped from two disconnect regions by merging temporally and spatially close segmentations.

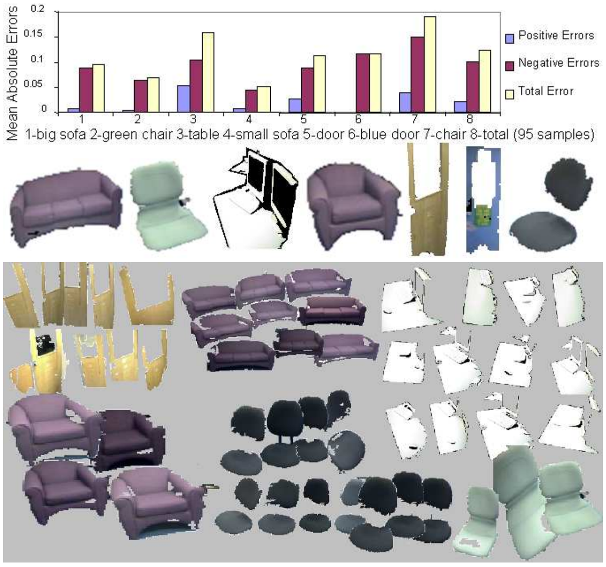


Figure 3. Statistics for furniture (random segmentation samples are shown). Errors given by (template area - object's real area)/(real area). Positive/negative errors stand for templates with larger/smaller area than the real area. Total stands for both errors.

Typical errors result from objects with similar color to their background, for which no perfect differentiation is possible, since the intersection of the object's compact cover of color regions with the object's complementary background is not empty. High color variability within an object create grouping difficulties (the compact cover contains too many sets – hard to group).

2.2. Attentional mechanism

Newborns have a special interest in oscillatory patterns of movements. During the first weeks of life, they focus attention on these type of movements for long periods of time. As previously described, we developed a mechanism that filters image data over time intervals according to its frequency content. However, this strategy only works if the human actor is able to engage the visual system, by having the active head gazing towards the object to be segmented.

An attentional Visual System [15] was therefore implemented to facilitate human-computer communication. This system combines salient stimulus from different feature modalities into a saliency map.The human actor gets visual attention to a desired object by creating a salient stimulus on such a target. The human waving behavior then primes the attentional system (such bias decreases with time) towards this stimulus (as shown in Figure 4).

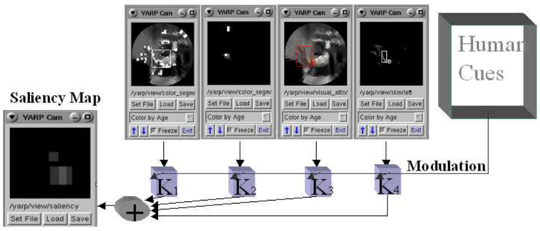


Figure 4. The attentional system running on the humanoid robot.

2.3. Object Recognition

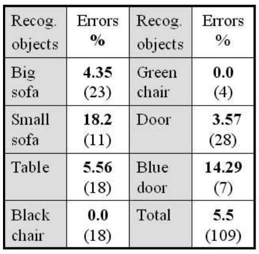
As just described, a human-computer interactive approach was implemented to introduce a humanoid robot to new percepts stored in its surrounding world. Such percepts are then converted into an useful format through an object recognition scheme, which enables the robot to recognize an object in several contexts and under different perspective views. This object recognition algorithm needs to cluster objects by classes according to their identity. Such task was implemented through color histograms – objects are classified based on the relative distribution of their color pixels (we developed recently a more elaborate object recognition algorithm, which processes independently chrominance, luminance and geometric cues [3]).

New object templates are classified according to their similarity with other object templates in an object database. A multi-target tracking algorithm (which tracks *good features* using the Lucas-Kanade Pyramidal algorithm) was developed to keep track of object locations as the visual percepts change due to movement of the active head. Table 1 presents performance statistics for this algorithm. It is also shown the system running on the humanoid robot.

3. 3D Environment Map Building

The world structure is a rich source of information for a visual system – even without visual feedback, people expect to find books on shelves. We argue that world structural information should be exploited in an active manner. For instance, there is a high probability of finding objects along the pointing direction of a human arm [12]. In addition, a human can be helpful for ambiguity removal: a human hand grabbing a Ferrari car implies that the latter is a toy car model, instead of a real car. Hence, humans can control the image context to facilitate the acquisition of percepts from a visual system.

We propose a real-time strategy to acquire depth information from monocular cues by having a human actor actively controlling the image context. It consists on automatically extracting the size of objects and their depth as a function of the human arm diameter. This diameter measure solves the image ambiguity between the depth and size of an object situated in the world.



Recog. objects	Errors %	Recog. objects	Errors %
Big sofa	**4.35** (23)	Green chair	**0.0** (4)
Small sofa	**18.2** (11)	Door	**3.57** (28)
Table	**5.56** (18)	Blue door	**14.29** (7)
Black chair	**0.0** (18)	Total	**5.5** (109)

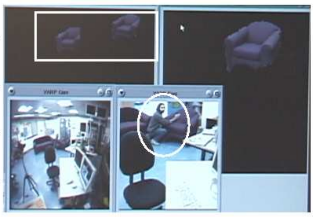


Table 1. (left) Recognition errors. It is shown the number of matches evaluated from a total of 11 scenes. Incorrect matches occurred due to color similarity among big/small sofas or between different objects. Missed matches result from drastic variations in light sources (right) object being recognized.

3.1. Coarse depth measures from Human Cues

Given the image of an object, its meaning is often a function of the surrounding context. The human arm diameter (which is assumed to remain approximately constant for the same depth, except for degenerate cases) is used as a reference for extracting relative depth information – without camera calibration. This measure is extracted from periodic signals of a human hand as follows:

1. Detection of skin-tone pixels over a image sequence
2. A blob detector labels these pixels into regions
3. These regions are tracked over the image sequence, and all non-periodic blobs are filtered out
4. A region filling algorithm (8-connectivity) extracts a mask for the arm
5. A color histogram is built for the background image. Points in the arm's mask having a large frequency on such histogram are labelled as background.
6. The smallest eigenvalue of the arm's mask gives an approximate measure of a fraction of the arm radius (templates shown in Figure 5).

Once a reference measure is available, it provides a coarse depth estimation in retinal coordinates for each arm's trajectory point. The following factors affect the depth estimation process (see Figure 6 for object reconstruction results, and Table 2 for an error analysis):

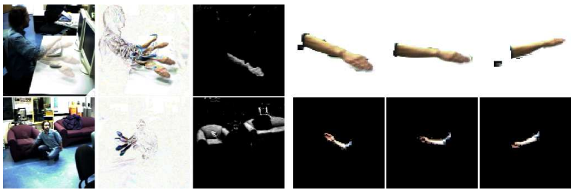


Figure 5. Human waving the arm to facilitate object segmentation. Upper row shows a sequence for which the skin-tone detector performs reasonably well under light saturation. Lower row shows background sofas with skin-like colors. The arm's reference size was manually measured as 5.83 pixels, while the estimated value was 5.70 pixels with standard deviation of 0.95 pixels.

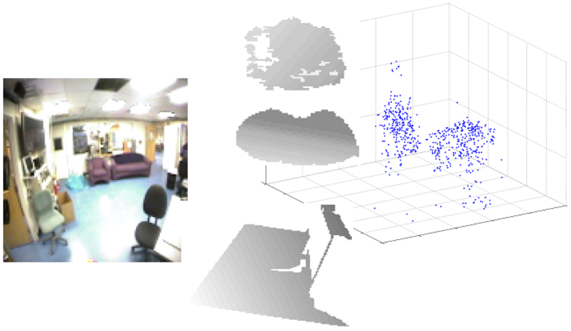
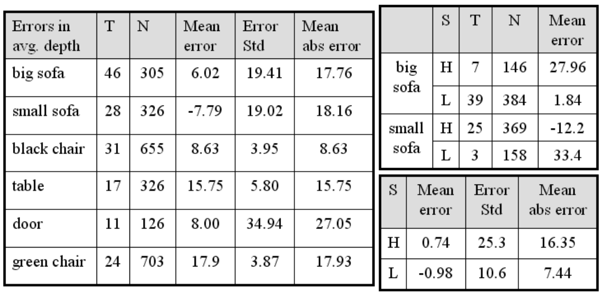


Figure 6. (left) An image of the lab. (right) Depth map (lighter=closer) for a table and a chair. Perpendicularity is preserved for the chair's disconnected regions (3D plot).

Light sensitivity This is mainly a limitation of the skin-color algorithm. We noticed a variation in between $10 - 25\%$ on size diameter for variations in light intensity (no a-priori environment setup – the only requirement concerns object visibility). High levels of light exposure increase average errors.

Human arm diameter variability Variations along people diversity are negligible if the same person describes objects in a scene to the visual system, while depth is extracted relative to that person's arm diameter.

Background texture interference The algorithm that we propose minimizes this disturbance by background removal. But in a worst case scenario of saturated, skin-color backgrounds, the largest variability detected for the arm's diameter was 35% larger than its real size.



Errors in avg. depth	T	N	Mean error	Error Std	Mean abs error
big sofa	46	305	6.02	19.41	17.76
small sofa	28	326	-7.79	19.02	18.16
black chair	31	655	8.63	3.95	8.63
table	17	326	15.75	5.80	15.75
door	11	126	8.00	34.94	27.05
green chair	24	703	17.9	3.87	17.93

	S	T	N	Mean error
big sofa	H	7	146	27.96
	L	39	384	1.84
small sofa	H	25	369	-12.2
	L	3	158	33.4

S	Mean error	Error Std	Mean abs error
H	0.74	25.3	16.35
L	-0.98	10.6	7.44

Table 2. Depth estimation errors for objects from 5 scenes (as percentage of real size). T **stands for number of templates,** N **for average number of trajectory points per template,** S **for light source,** H **and** L **for High/Low luminosity levels, respectively. (left) overal results (right) Depth errors for different luminosity conditions are shown for the two sofas – top – and from all objects– bottom.**

Hence, we argue that this technique provides coarse depth estimates, instead of precise, accurate ones. The average depth of an object can be estimated by averaging measures using a least squares minimization criterium – errors are even further reduced if large trajectories are available.

But since a collection of 3D trajectories (2-dimensional positions and depth) are available from temporally and spatially closed segmentations, it is also possible to determine a coarse estimate for the shape of an object from such data. A plane is fitted (in the least square sense) to the 3D data, for each connected region in the object's template – although hyperplanes of higher dimensions or even splines could be used. Outliers are removed by imposing upper bounds on the distance from each point to the plane (normalized by the data standard deviation). Such fitting depends significantly on the area covered by the arm's trajectory and the amount of available data. The estimation problem is ill-conditioned if not enough trajectory points are extracted along one object's eigendirection. Therefore, the fitting estimation depends on the human description of the object – accuracy increases with the area span by the human trajectories and the number of extracted trajectory points.

3.2. Map Building from Human Contextual Cues

The object location $p = (\theta, \psi)$ in the active vision head's gazing angles (egocentric coordinates), together with the estimated depth and the object's size and orientation, are saved for further processing. Each point in the object's template is converted to egocentric coordinates using a motor-retinal map (obtained by locally weighted regression).

A scene was defined as a collection of objects with an uncertain geometric configuration, each object being within a minimum distance from at least one other object in the scene. Figure 7 presents both coarse depth images and 3D reconstruction data for a typical scene in the robot's lab. The geometry of a scene was reconstructed from the egocentric coordinates of all points lying on the most recent object's template. Figure 8 presents further scene reconstruction results without deformation.

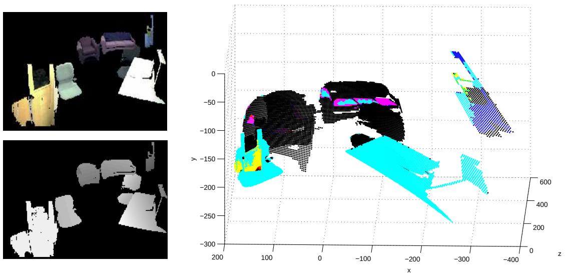


Figure 7. (left) Furniture image segmentations – on top – and depth map – bottom – for the scene in Figure 8; (right) Coarse 3D map of the same scene. Depth is represented on the axis pointing inside, while the two other axis correspond to egocentric gazing angles (and hence the spherical deformation).

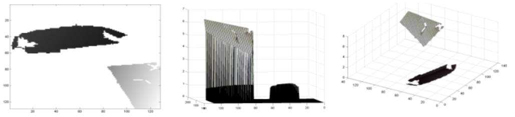


Figure 8. (left) Depth map for a table and a sofa (right) two views of the 3D reconstruction.

Scene reconstruction was evaluated from a set of 11 scenes built from human cues, with an average of 4.8 objects per scene (from a set of ten different furniture items). Seven of such scenes were reconstructed with no object recognition error, and hence for such cases the scene organization was recovered without structural errors. An average of 0.45 object recognition errors occurred per scene.

4. Conclusions

This paper presented an alternative strategy to extract depth information. The method proposed relies on a human actor to modify image context so that percepts are easily perceived – a waving human arm in front of an object provides an important cue concerning the size of such object.

Throughout this paper percepts were acquired by an active vision head on a stationary platform (the humanoid robot). This work is being extended to a mobile platform, for performing simultaneously map building and robot localization. Whenever a scene object is recognized, the system actively searches for other objects. If more than one object appear displaced, then all objects are used as natural landmarks for robot localization. Otherwise, the scene 3D model and the spatial distribution of objects are updated.

This human-centered framework is also being currently applied to teach robots from books [3]; to generate training data for contextual priming of the attentional focus from holistic cues; to learn cross-modal properties of objects, by correlating periodic visual events with periodic acoustic signals; and there is still a hough number of potential applications for which this approach might bring benefits.

References

[1] J. Aloimonos, I. Weiss, and A. Bandopadhay. Active vision. *Int. Journal on Computer Vision*, 2:333–356, 1987.

[2] M. Anderson. Embodied cognition: A field guide. *Artificial Intelligence*, pages 91–130, 2003.

[3] author. Teaching a humanoid robot from books. 2004.

[4] author. Towards and embodied and situated ai. 2004.

[5] R. Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):996–1005, August 1988.

[6] R. Chatila and J. Laumond. *Position referencing and consistent world modelling for mobile robots*. IEEE International Conference on Robotics and Automation, 1985.

[7] D. Comaniciu and P. Meer. Robust analysis of feature spaces: Color image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, San Juan, 1997.

[8] J. Harris. *Algebraic Geometry: A First Course (Graduate Texts in Mathematics, 133)*. Springer-Verlag, January 1994.

[9] E. Krotkov, K. Henriksen, and R. Kories. Stereo ranging from verging cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(12):1200–1205, 1990.

[10] Y. Kuniyoshi, M. Inaba, and H. Inoue. Learning by watching: Extracting reusable task knowledge from visual observation of human performance. *IEEE Transactions on Robotics and Automation*, 6(10), 1994.

[11] M. Nicolescu and M. Mataric. Experience-based learning of task representations from human-robot interaction. In *IEEE International CIRA Symposium*, 2001.

[12] D. I. Perrett, A. J. Mistlin, M. H. Harries, and A. J. Chitty. Understanding the visual appearance and consequence of hand action. In *Vision and action: the control of grasping*, pages 163–180. Ablex, Norwood, NJ, 1990.

[13] V. Sequeira. *Active Range Sensing for Three-Dimensional Environment Reconstruction*. PhD thesis, Department of Electrical and Computer Engineering, IST/UTL, 1996.

[14] A. Torralba and A. Oliva. *Global depth perception from familiar scene structure*. MIT AI-Memo 2001-036, CBCL Memo 213, December 2001.

[15] J. M. Wolfe. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, 1(2):202–238, 1994.