Attention-Based Target Tracking for an Augmented Reality Application

Morgan Veyret
European Center for Virtual Reality, France
veyret@enib.fr

Eric Maisel
European Center for Virtual Reality, France
maisel@enib.fr

ABSTRACT

When visiting an aquarium, people may be disturbed or, at least, disappointed by the amount and diversity of available information. Moreover, one can find it very difficult to match the information of notices on the wall to the reality of the fishes. Therefore, we propose a virtual guide, an autonomous teaching assistant embodied in the real world using augmented reality techniques, for helping people in their visit of aquariums. This virtual guide will interact with the real world using multiple modalities (e.g. speech, facial expression, ...). Thus, it should be aware of the aquarium's state and content, and use perceived information and prior knowledge to inform the visitor in a structured fashion. Due to the high mobility and unpredictable behaviour of the fishes, our guide requires an adequate perception system. This camera-based system has to keep track of the fishes and their behavior. It is based on the focalisation of visual attention that allows to select interesting information in the field of view. This is achieved by extracting a number of focuses of attention (FOA) using a saliency map and a multi-level memory system, which is filled (or updated) with the extracted information. It allows our system to detect and track targets in the aquarium. This article describes how we use the saliency map and memory system, along with their interactions, to set up the first part of our perception system.

Keywords: Augmented Reality, Virtual Guide, Autonomous, Perception, Computer Vision, Tracking, Saliency Map, Memory

1 INTRODUCTION

When visiting an aquarium, matching notices on the wall with the content of the aquarium can be very difficult. Even if one can find the appropriate notice for a specific fish, when looking back to the aquarium, this fish may have moved or be hidden. The attractiveness and pedagogical goal of the aquarium may suffer from these problems.

To solve such a problem, we propose to use a virtual guide to help people during their visit of an aquarium. This guide will be embodied in the real world using augmented reality techniques [1] and will interact with its environment (see figure 1). It will use several modalities to deliver information concerning the aquarium and its contents to the visitor like speech and facial expressions.

The guide's discourse will have to be structured based on both prior static knowledge concerning the aquarium (and the fishes it contains) and dynamic knowledge of the activity inside the aquarium. To build such a discourse, our guide will have to extract information related to its current concerns. As an example, while the guide is talking about a shark, it may be more interesting to find information about other shark or the shark's behavior. On the opposite, if its talk is about to end and another fish, that hasn't been described yet, appears, the guide may use this opportunity to start talking about another topic. The above situations show the need for a specific perception system that can be controled by our guide while providing information about unexpected events.

Figure 1: The virtual guide embodied in the real world to help people during their visit.

We plan to build such a perception system using the focalisation of visual attention in a hybrid model. This model will take into account the two main influences of visual attention (see section 2). Using such an approach, our perception system may show the control and reactivity compromise our guide requires.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SHORT COMMUNICATION proceedings
ISBN 80-86943-05-4
WSCG'2006, January 30 – February 3, 2006
Plzen, Czech Republic.
Copyright UNION Agency – Science Press

[1] Projection on a transparent screen

This article describes the first part of the perception system that is able to detect and track targets using the focalisation of attention to select interesting information out of video streams. This is achieved using a simplified classic attention model [30, 18, 14] based on the use of a saliency map in addition to a memory system. The saliency map is used to extract focuses of attention (FOA) according to the encoded importance of visual features. Those FOAs are then used as an input for a multi-level memory system, which is used to store information about targets over time and track them. Here we show how our model works and may be used in the planned guide 's perception system based on a hybrid visual attention model.

This document is organized as follow: the first section of this article will describe a few domains our model is connected to, which includes the focalisation of visual attention, as well as target tracking models. The second section describes our work in details while the third shows some results concerning target tracking using the proposed approach on aquarium static [2] videos. In the last section, we will consider what remains to be done for the virtual guide application and the perception model in particular.

2 RELATED WORK

A classic model of visual attention is proposed by Koch and Ullman in 1985 [18]. Their architecture describes the attentional process based on two steps: a parallel process followed by a sequential one. The first step is based on features maps computed over the entire field of view and combined in a subsequent saliency map (see figure 2). Then a Winner-Take-All (WTA) algorithm is used to extract the most salient locations in a sequential manner. An *inhibition of return* (IOR) mecanism is set up to avoid multiple selections of the same location. They considered a small number of basic features that have been found to be used in the human visual system like: colors, orientations and contrast. In 1998, Itti *et al.* [14] proposed a similar model of bottom-up (BU) visual attention that showed good results at simulating human focalisation of visual attention. They used center-surround feature computation and a new combination strategy. Later, they proposed a new combination strategy [12] to improve their saliency map model. The proposed strategy is designed to enhance locations that differ from their surrounding. Other BU perception models were proposed: Terzopoulos *et al.* [28] proposed a simulation of virtual fishes and their environment. They used peripheral vision through a number of concentric virtual cameras to model the focus of attention. In [8], Courty and Marchand proposed to use simplified saliency maps

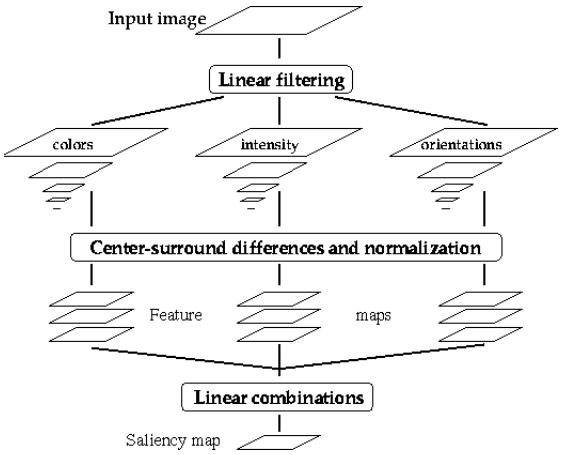

Figure 2: Itti *et al.* saliency map model. Features are extracted and combined across different scales in a center-surround manner. The obtained feature maps are linearly combined to build the saliency map.

to model visual attention and improve the realism of their virtual actor. This was done by making the actor look towards salient location (in 3D space). In such BU models, salient locations are areas that correspond to the most active features considered. This allows a non-negligable reactivity during the perception process. Those classic models describe the focalisation of visual attention in a bottom-up manner and are closely related to Treisman's *Feature Integration Theory* [30] and Wolfe's *Guided Search* [33]. However, Wolfe's model proposed top-down (TD) cueing of the feature maps that used prior knowledge in active search tasks. For example,when looking for a red object, a color feature map can be biased to account only for red color, improving the saliency of red objects in the field of view. In 2002, Itti and Navalpakkam [22] proposed a new visual attention model based on Itti's earlier work [14, 15, 13] integrating TD influence and a memory system. This influence biased the saliency map through the modulation by a so-called "attention map" which was built using prior knowledge contained in memory and concerning the search task as well as recently known information (from a working memory). In 2005 [23], they improved their model with the noticable addition of recognition. Olivia *et al.* [25] also used the notion of saliency map (defined as a probability of feature presence) and its modulation to build an attention model reflecting both BU and TD path. These hybrid approachs take advantages from BU models while enabling active search tasks which requires the control of visual attention.

Hayhoe *et al.* [11] also showed the presence of control in visual attention as well as different memory level ranging from short term (working) memory to an "unlimited" long term one. Several perception systems used memory systems. Kuffner and Latombe [20] used

[2] Camera's position is fixed

it to store information about perceived objects (position, last time seen, ...) and then used it in an inhibition of return mecanism, which was based on objects rather than classic spatial locations. Noser *et al.* [24] used the notion of visual memory and synthetic vision to model the perception of a digital actor which used this memory to find a path through its environment in an obstacle avoidance problem.

Work have also been done concerning target tracking, mainly for video based surveillance systems [6], robots vision [7] and augmented reality registration [17]. A classic approach is to extract feature points (e.g. markers or corners) and track those points over time [29, 31] based on trajectory and movement asumptions to help reducing matching possibilities. These approachs are known as *Feature Based Tracking*. Burghardt *et al.* used this approach to detect and track animal faces in wildlife footages [3]. Chetverikov and Verestoy [5] used a similar approach to track dense feature point sets. Feature tracking was also used by Coifman *et al.* [6] to detect and track moving vehicles making multiple assumptions like fixed cameras and car's straight trajectory. Other authors used an active contours approach [27] to track 3D objects' pose in 2D space. Another approach is to try to match an area of a frame with a known model [19, 4, 7] (e.g. color distribution [2]). This approach is known as *Model Based Tracking* (or *Region Based Tracking* [21]). Ramanan and Forsyth used such an approach in 2003 [26] to detect and track a human based on a model of its body parts, using bayesian network inference. There was also recent attemps to build a tracking system based on the focalisation of attention [10] to help oceanographic researcher to annotate underwater video sequences.

3 OUR WORK

The proposed model is inspired from biological and psychological studies about visual attention [30, 11], work in computer vision concerning saliency maps [14] and the focalisation of visual attention in general.

3.1 Bottom-up Focalisation of Visual Attention

As we said in the introduction of this document, we want to build a perception system for a virtual guide. This perception system must be able to provide unexpected information as well as requested one. We are looking at creating a tracking system based on visual attention to allow the future guide perception behavior (using an hybrid visual attention model integrating both low-level and high-level influences). Here we only describe a "Bottom-Up" approach that is used to detect and track targets in the aquarium.

Saliency Map Our tracking system is inspired from work by Itti *et al.* [14] on saliency maps (SM). They built a SM model (see figure 2), based on biologically

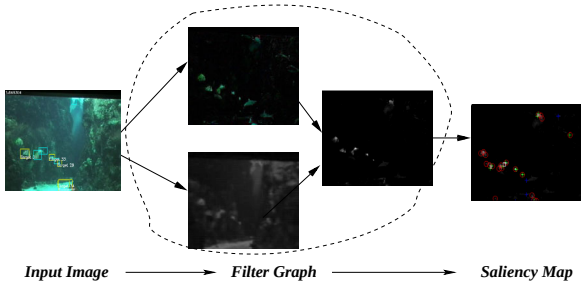

Figure 3: Saliency map creation process: an input image (on the left) is passed throught a filter graph (in the middle) representing the creation process and the saliency map is outputed (on the right)

inspired features maps, that showed good results at predicting focus of attention localisation when facing natural and noisy images. Their system was able to predict most of the salient locations a human visual system would have looked at. The saliency map is responsible of encoding the interest of image areas according to considered features (intensity, colors, orientations,...). In [32], we used a simplified saliency map model using biologically inspired features to predict salient locations in a visual attention system designed for understanding road situations. In this model, we proposed a loop where saliency was used to modify entries of a bayesian network and the output of this network was used to modify the saliency map creation process. We wanted the saliency computation system designed for this work to be modular in order to allow modifications of the saliency map creation process (to model the top-down influence through features biasing).

This system is based on the notion of *filter*. A filter is an object that handles image processing operations. Each filter may have a set of parameters that allows us to control its behavior as well as a number of inputs and outputs. Filters may be connected to each other in a directed acyclic graph (DAG) to describe the entire saliency map creation process (see figure 4 and 3). At the top of this graph, there is the input image for which we want to compute saliency and the output node correspond to the computed saliency map.

Different kinds of filters are available like background substraction operator, color filtering (Red, Green, Blue, Yellow) or multiscale pyramid creation. Other filters are combination operators that allow different combination strategies between feature maps like maximum combination (where the maximum value of each input map is kept in the resulting map) or weighted combination.

Using this modular scheme, we are able to build the desired saliency map by describing the filtering sequence as a graph. Resulting map is computed in a backward manner, meaning that each filter requests its

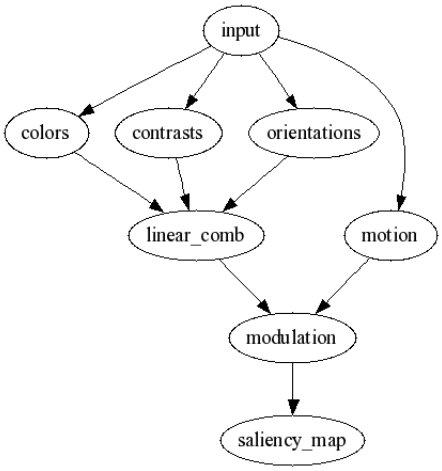

Figure 4: Example of a directed acyclic graph (DAG) that represent one saliency map creation process

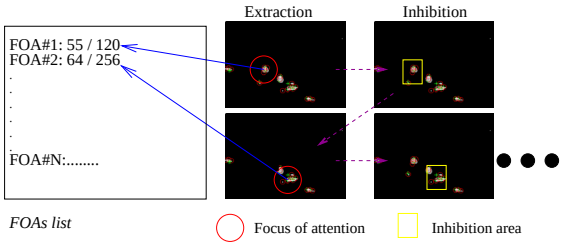

Figure 5: Extraction of focuses of attention with inhibition of return (IOR). The algorithm output a requested number of points by repeating an, extraction and inhibition process.

inputs from its parents in the graph. Then, only required computation is done.

For now, we use a simplified version of saliency maps in order to allow for real-time computations. Hence we only use a limited amount of features (colors, motion, maybe orientations) that are linearly combined into the resulting saliency map. Our modular architecture allows us to experiment multiple saliency map models.

Focalisation Once the saliency map (SM) is computed, it can be used to extract multiple focuses of attention. A focus of attention (FOA) is an area of the field of view corresponding to the most salient location in the saliency map. This focus of attention is defined by a position and a size in the saliency map coordinate system (in pixels). FOAs are extracted from the saliency map by using a Winner-Take-All (WTA) style algorithm. We tryied different extraction algorithms. Since we want to extract multiple FOAs, we set up an "inhibition of return" mecanism (like in [14, 9]) to ensure that the same location is not selected multiple times (see figure 5).

3.2 Memory System

The bottom-up attention system described above doesn't store information over time. Therefore it is not sufficient to set up the requested tracking system. To allow to store perception information from the saliency map over time, we propose a memory system that will enable tracking.

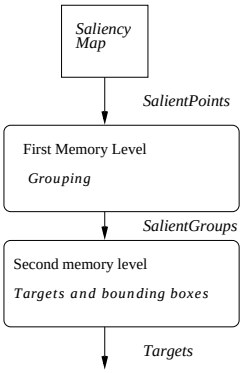

Figure 6: Overview of the different memory levels and the objects they handle.

This memory system is composed of two main levels.The first level, which can be considered as a short term (or working) memory, stores information about focuses of attention, while the second one, which can be considered as a medium term memory, uses information from the short term memory to detect and track targets over time (see figure 6).

First Level The first memory level is based on *salient points* which are built from the extracted focuses of attention. A salient point associates multiple informations to a location on the saliency map (FOA). As an example, it may contain information about intermediate saliency values. These salient points are the base elements of our memory system and are used to fill its first level.

For each of this points presented to the memory input, we try to match an existing one (already stored in this memory level) by using an euclidian distance computed in image space (feature space computation are also planned). If the computed distance is less than a specified threshold (FOA size in the case of image space distance) then points are considered to match each other.

To enhance this simple matching process, we use Kalman's filtering techniques to predict positions of existing points in memory and distance is computed using the predicted point's position.

If a point is matched correctly with an existing one, we update existing information with the newly perceived point. This update procedure means correcting the Kalman's filter parameters to take into account the perceived trajectory modifications as well as updating lifetime properties of the point in memory

Here, each point has a timestamp, indicating its last update time, and other time properties: maximum lifetime, active time and time spent in memory.

Maximum lifetime is used to determine wether a point has to be removed from memory or not. As this level is "short term", we use a small lifetime value (generally < 3 seconds). The active time parameter serves as an up-to-date marker that allows us to know if the point has been updated recently (this value is smaller than the maximum lifetime of a point since we consider that the point is up-to-date only a few instants after being updated). The last parameter is a focalisation one that enables us to consider only points that have been in memory for a sufficiently long time.

If the point isn't matched with an existing one, then it is added in memory. There's a limitation concerning the number of points that can be present simultaneously in the memory. This limitation allows us to add a new level of focalisation by considering only points that are perceived multiple times (this also avoids some possible noise from the saliency map).

Once we have updated our memory with new perceived FOAs, active points (recently perceived) that have been there for sufficient time are grouped to take into account spatial "similarities". Thus, considering the salient point sizes, we build groups according to their distance in the image [3]. This allows us to reduce the computation time required to process those points.

Grouping is also inspired by the human visual system where groups are built during perception (e.g. based on features or spatial similarities).

Those groups will serve as an input for the second memory level (medium term memory).

Second Level At this level, we manipulate groups of salient points and *targets*.

A target describes an area of the processed image that may be considered as an object. It stores information about this area as a bounding box defined by its size and position.

For each group, extracted from the previous memory level, we check if an associated target exists. If such a target is found, then it is updated directly from the group without any further test. If there is no associated target, then a new target is created.

Each target uses a Kalman's filter to predict both its size and position from one frame to the next. When updating a target, lifetime and Kalman's filter parameters are modified according to the perceived information. To extract the bounding box's size and position from the salient group, we use a blob extraction algorithm based on thresholding that allows us to extract size limits of a segmented image area. If we can't find bounding box information using this segmentation algorithm (due to bad segmentation or thresholding problem for example), we use the salient group information to extract an approximation of the expected bounding box. This is done by computing minimum and maximum coordinates, using the points in the associated group, and using this information as the bouding box. If the target has been updated before, this won't affect the Kalman's filter too much thanks to the previously perceived information.

Lifetime management is the same as for the preceding memory level except that the targets have no indication of time spent in memory. Moreover, this level doesn't have any capacity limitation. This was not necessary since focalisation is done at the preceding memory level.

This memory system and the interactions between the different levels set up a new focalisation mecanism on top of the saliency map we used to extract focus of attention. Our system is then able to focalize on targets that are already known for a long time while allowing unexpected targets perception, due to the use of the saliency map. This is the first part of our perception system where targets will be used by our guide to build a representation of the real world. Then it will use this representation to build an explanation and interact with the visitors.

[3] We also plan to consider feature space to check for similarities between points

4 EXPERIMENTS

We implemented our system using the OpenCV [1] computer vision library that provides low-level image processing functions as well as higher level algorithms for computer vision applications.

To test the system, we used a video of a tropical aquarium recorded at a 320x240 pixels resolution, shot with a digital video camera. The video was played in a different thread in our tracking system at 25 FPS, which is the expected frame rate from the camera system we envisage. Our system grabs frames while the video is still being played at the same frame rate. This is why we use the *number of processed frames* as a time measurement in our results.

We used a saliency map based on simple features: motion, intensity and colors. To extract the motion mask, we used the algorithm from KaewTraKulPong and Bowden [16] which is implemented in the OpenCV [1] library. Other features are extracted as follow:

- Intensity is extracted directly out of color corrected images by computing mean of all channels:

$$I(x,y) = \frac{R(x,y)+G(x,y)+B(x,y)}{3}$$

- Colors are extracted using Itti's [14] formulaes:

$$R = r - \frac{g+b}{2} \qquad G = g - \frac{r+b}{2}$$

$$B = b - \frac{r+g}{2} \qquad Y = \frac{r+g}{2} - \frac{|r-g|}{2}$$

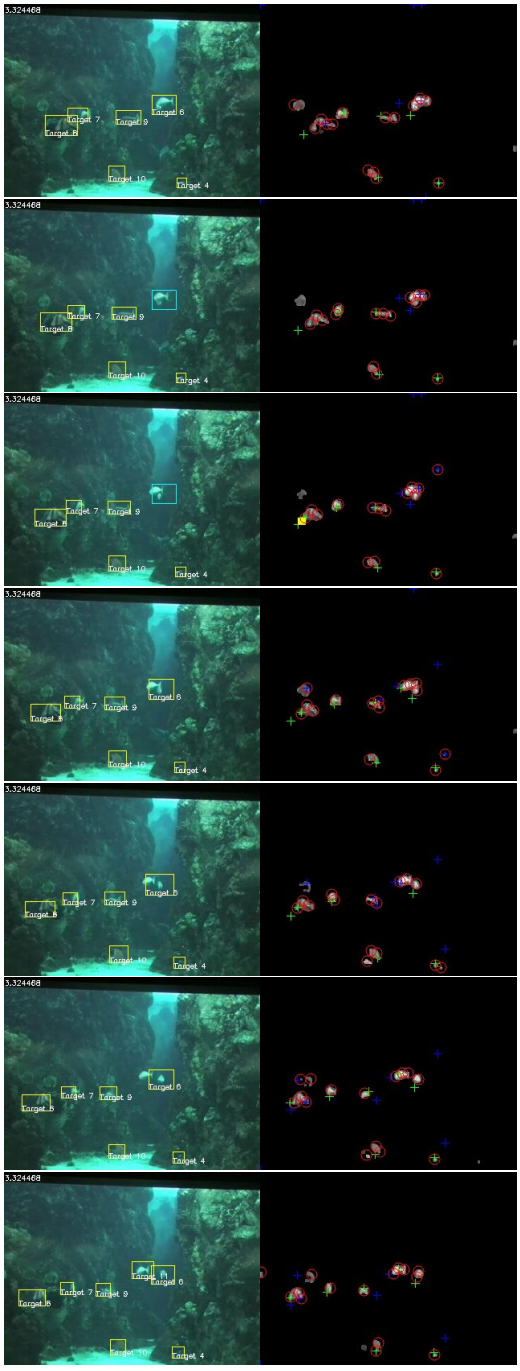

Figure 7: Tracking sequence with corresponding saliency maps (from top to bottom). The left column shows tracking output with yellow rectangle to represent currently tracked targets and light blue ones for lost targets that still being in memory. The right column show the corresponding saliency maps with the current focuses of attention as red circle, salient point in memory as cross (green for active one and blue for those we're forgetting) and salient groups as filled yellow rectangles.

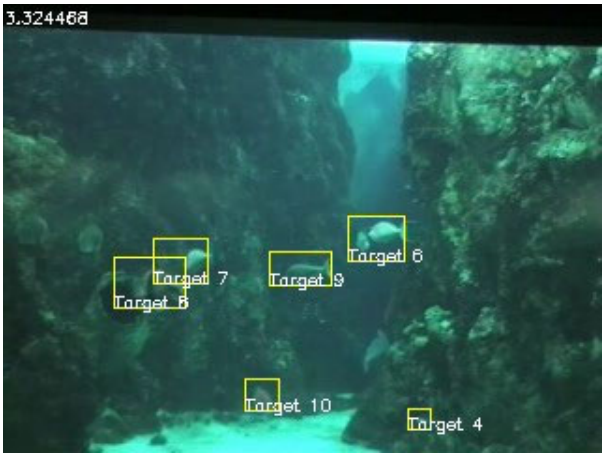

Figure 8: Tracking results. "Target 4" is a false positive due to some illumination artifacts.

Simple feature maps (intensity and colors) are combined into a global map using a maximum combination operator:

$$Global(x,y) = \max_{i \in \mathbb{F}}(F_i(x,y)).$$

Then the motion map is used as a modulation parameter to obtain the final saliency map:

$$SM(x,y) = Global(x,y) \bullet Motion(x,y)$$

There was no need of a training phase and our system was able to detect and track targets at a rate of about 5 FPS, which is acceptable for our real-time purpose since no real optimisation has be done yet.

Figure 7 shows the tracking process over a few number of consecutive processed frames. We can see that our system is able to detect and track multiple targets at the same time (this may be partially controled by the number and size of extracted focuses of attention). Targets may also be lost for a certain number of frames. When lost, a target can be recovered if updated by a new salient point thanks to our memory system that keeps track of targets during a certain amount of time.

Focuses of attention, salient points and groups can be seen on figure 9. Groups help our system when a previously merged target (two overlapping objects) is splited. The apparition of two groups instead of the previous one allows us to create a new target to track the newly discovered object.

We can also notice the presence of a false target detection (see figure 8) that passed successfully through our memory system. This target is detected due to some illumination artifacts that persisted. This accounts for the need of a control system. Recognition, or trajectories information for example, can be used to disable this kind of targets and enhance the perception system.

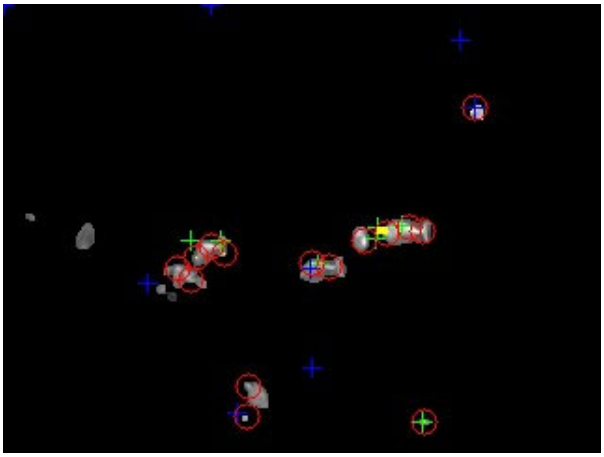

Figure 9: Saliency map. FOAs are the red circles, salient points are represented as crosses (active ones are in green) and salient groups as yellow filled rectangles.

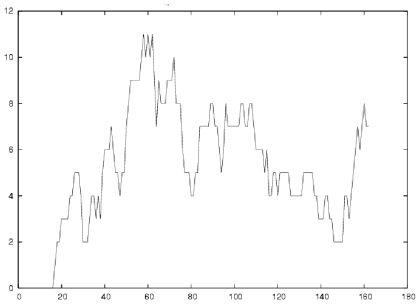

Figure 10: Evolution of the number of targets (vertical axis) over time (horizontal axis), time is defined in number of processed frames

We also measured the number of targets that are tracked at once. As showed in figure 10, this number is limited even though our second memory level has no capacity limitation. This is due to the capacity of the first memory level that keeps our system from accounting for every single salient location detected. We can also see the memory "initialisation process" through the absence of tracked targets at the beginning of the experiment. This is due to the time needed by salient points to be allowed to go to the second memory level.

Figure 11 shows the lifetime of some tracked targets in terms of number of processed frames[4]. This lifetime ranges from about 4 processed frames to 60. The average is about 13, thus showing that our simple tracking system is able to track targets over time without using complex trajectory models or assumptions. For now, new targets aren't matched with dead ones as they aren't stored in memory. This would require a long term

[4] A processed frame is different from consecutive frame since the camera can grab multiple consecutive frames while only one frame is processed

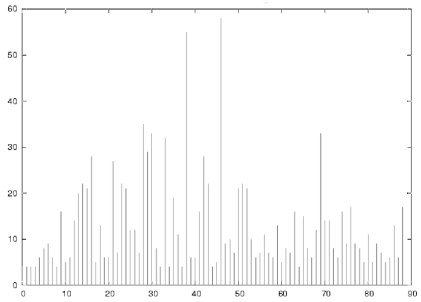

Figure 11: Lifetime of targets in number of processed frames (vertical axis), each impulse correspond to a target

memory storing information on old targets and is more appropriate for attention control. This long term memory could, for example, use trajectories to try to match new targets with dead ones and, thus, improving tracking capabilities.

5 DISCUSSION AND FUTURE WORK

The tracking system presented here is only a part of our proposed guide perception system which needs to include attentional control. This leads to the necessity of a recognition system to identify tracked fishes and the modelling of our guide's goals. These goals will then be used to control the focalisation of visual attention. We'll probably need to define visual strategies to abstract this control. We'll also need to describe our guide's knowledge in order to use it in the perception process.

The limited size of the tracked targets will also be a serious limitation when trying to recognize the fishes[5]. To overcome this difficulty, we plan to use an extended camera system with two movable cameras in addition to the fixed one used for tracking. Those cameras will have to focus on selected targets using information provided by our tracking system.

The next step will consist in adding recognition capabilities to our system and use it in our guide perception system. The intented recognition approach will be based on previous work [32]. Based on focused targets (from the extended camera system), we plan to extract specific salient information (e.g. a tuple containing a color, its spatial location on the target and a saliency value) and use this information to update a belief network. This network will then be used in a backward manner to select the information to extract.

6 CONCLUSION

We proposed a virtual guide to help people in their visit of aquariums. We described the needs of perception of

[5] Average target size is about twenty pixels large.

this guide and presented a new tracking system based on the focalisation of visual attention.

This system used a saliency map inspired from a biological model in association with a multi-level memory. It showed some tracking capabilities when tested against static (no camera movement) videos of an aquarium. Fishes were succesfully detected and tracked over a few frames while the total number of considered information was still low.

We have also given a brief overview of the intented recognition process and work that still need to be done in order to achieve the virtual guide.

REFERENCES

[1] Opencv computer vision library. http://www.intel.com/research/mrl/research/opencv/.

[2] Gary R. Bradski. Computer vision face tracking for use in a perceptual user interface. *Intel Technology Journal*, 2(2):1–15, 1998.

[3] Tilo Burghardt, Janko Calic, and Barry Thomas. Tracking animals in wildlife videos using face detection. In *European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology*, October 2004.

[4] Andrea Cavallaro, Olivier Steiger, and Touradj Ebrahimi. Tracking video ob jects in cluttered background. In *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY*, 2004.

[5] D. Chetverikov and J. Verestoy. Motion tracking of dense feature point sets. In *Proc. 21th Workshop of the Austrian Pattern Recognition Group*, pages 233–242, Halstatt, Oldenbourg Verlag, 1997.

[6] B. Coifman, D. Beymer, P. McLauchlan, and J. Malik. A real-time computer vision system for vehicle tracking and traffic surveillance. *Transportation Research: Part C*, 6(4):271–288, 1998.

[7] Andrew I. Comport, Éric Marchand, and François Chaumette. Robust model-based tracking for robot vision. In *IEEE/RSJ Int. Conf on Inteligent Robots and Systems, IROS'04*, Sendai, Japan, September 2004.

[8] Nicolas Courty and Eric Marchand. Visual perception based on salient features. In *Proceedings of the 2003 IEEE/RSJ Intl. Conference on Intelligent Robots and Systems*, Las Vegas, Nevada, October 2003.

[9] Nicolas Courty, Eric Marchand, and Bruno Arnaldi. A new application for saliency maps: Synthetic vision of autonomous actors. In *IEEE Int. Conf. on Image Processing, ICIP'03 Barcelona, Spain*, September 2003.

[10] Walther Dirk, Duane R. Edgington, and Christof Koch. Detection and tracking of objects in underwater video. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Washington, D.C., 2004.

[11] Mary M. Hayhoe, Dana H. Ballard, Hiroyuki Shinoda Jochen J. Triesch, Pilar Aivar, and Brain T. Sullivan. Vision in natural and virtual environments. *Eye Tracking Reasearch and Applications Symposium*, 2002.

[12] L. Itti and C. Koch. A comparison of feature combination strategies for saliency-based visual attention systems. In *SPIE Human Vision and Electronic Imaging (HVEI'99), San José, CA*, pages 373–382, January 1999.

[13] L. Itti and C. Koch. A saliency based search mechanism for overt and covert shifts of visual attention. In *Vision Reasearch **40***, pages 1489–1506, May 2000.

[14] L. Itti, C. Koch, and E. Niebur. A model of salyency-based visual attention for rapid scene analysis. In *IEEE Transactions on Pattern Analysis and Machine Intelligence **20***, pages 1254–1259, November 1998.

[15] Laurent Itti and Christof Koch. Learning to detect salient objects in natural scenes using visual attention. In *Image Understanding Workshop*, 1999. (in press).

[16] P. KaewTraKulPong and R.Bowden. An improved adaptive background mixture model for real-time tracking with shadow detection. In *2nd European Workshop on Advanced Video Based Surveillance Systems*, 2001.

[17] Georg Klein and Tom Drummond. Robust visual tracking for non-instrumented augmented reality. In *The Second IEEE and ACM International Symposium on Mixed and Augmented Reality*, page 113, 2003.

[18] C. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, January 1985.

[19] D. Koller, K. Daniilidis, and H. H. Nagel. Model-based object tracking in monocular image sequence of road traffic scenes. *International Journal of Computer Vision*, 10:257–281, 1993.

[20] J. Kuffner and J. Latombe. Fast synthetic vision, memory, and learning models for virtual humans. *Computer Animation*, pages 118–127, 1999.

[21] François Meyer and Patrick Bouthemy. Region-based tracking in an image sequence. Technical report, INRIA (Programme 4) : Robotique, Image et Vision, 1992.

[22] V. Navalpakkam and L. Itti. A goal oriented attention guidance model. In *Proc. 2nd Workshop on Biollogically Motivated Computer Vision (BMCV'02)*, pages 453–461, Tuebingen, Germany, November 2002.

[23] V. Navalpakkam and L. Itti. Modeling the influence of task on attention. *Vision Research*, 45:205–231, 2005.

[24] H. Noser, O. Renault, and D. Thalman. Navigation for digital actors based on synthetic vision, memory and learning. *Computer & Graphics*, 19(19):7–19, 1995.

[25] Aude Olivia, Antonio Torralba, Monica S. Castelhano, and John M. Henderson. Top-down control of visual attention in object detection. In *Proceedings of the IEEE International Conference on Image Processing*, volume 1, pages 253–256, Barcelona, Spain, September 2003.

[26] D. Ramanan and D. A. Forsyth. Finding and tracking people from the bottom up. In *Computer Vision and Pattern Recognition (CVPR)*, Madison, WI, June 2003.

[27] K. Stark. A method for tracking the pose of known 3-d objects based on an active contour model. In *ICPR96*, pages 905–909, Vienna, August 1996.

[28] Demetri Terzopoulos, Tamer Rabie, and Radek Grzesczuck. Perception and learning in artificial animals. In *Artificial Life V : Proc. Fifth Inter. Conf. on the Synthesis and Simulation of Living Systems, Nara, Japan*, May 1996.

[29] Carlos Tomasi and Takeo Kanade. Detection and tracking of point features. Technical report, Carnegie Mellon University, April 1991. CMU-CS-91-132.

[30] Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, January 1980.

[31] J. Verestoy and D. Chetverikov. Comparative performance evaluation of four feature point tracking techniques. In *Proc. 22nd Workshop of the Austrian Pattern Recognition Group*, pages 255–263, Illmitz, Austria, 1998.

[32] Morgan Veyret and Eric Maisel. Simulation de la focalisation de l'attention visuelle: application à la simulation d'automobilistes virtuels. In *AFIG*, Poitiers, 2004.

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