Unified Visual Perception Model for Context-aware Wearable AR

Youngkyoon Jang†
PhD Candidate

Woontack Woo‡
Thesis Supervisor

UVR Lab, Graduate School of Culture Technology, Korea Advanced Institute of Science and Technology

Abstract

We propose Unified Visual Perception Model (UVPM), which imitates the human visual perception process, for the stable object recognition necessarily required for augmented reality (AR) in the field. The proposed model is designed based on the theoretical bases in the field of cognitive informatics, brain research and psychological science. The proposed model consists of Working Memory (WM) in charge of low-level processing (in a bottom-up manner), Long-Term Memory (LTM) and Short-Term Memory (STM), which are in charge of high-level processing (in a top-down manner). WM and LTM/STM are mutually complementary to increase recognition accuracies. By implementing the initial prototype of each boxes of the model, we could know that the proposed model works for stable object recognition. The proposed model is available to support context-aware AR with the optical see-through HMD.

Index Terms: H.5.1 [Information Interfaces and Representation (HCI)]: Multimedia Information Systems—Artificial, augmented, and virtual realities; I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Object recognition

1 Introduction

According to the advances of optical see-through HMD and the advent of the recent glasses-type devices, there are rapidly increasing demands on AR technology, which moves forward to a "practical field application" considered in real situations. To meet the requirements, the first stage for implementing AR must be a stable recognition of target objects or a space within a captured scene image. For stable recognition, there have been studies regarding computationally optimized approaches according to different targets or based on different devices. In other words, through the advances of hardware performance and devices, recent approaches to computer vision have overcome various problems in many different aspects of computer vision research area that were previously difficult to solve.

As mentioned, there have been tremendous advances in the approaches, especially for object recognition based on the advances of hardware and devices. For example, according to the advent of RGBD camera, [21, 12, 14] solve real-time three-dimensional scene registration problems at space level, while [15, 24] solve three-dimensional object detection and tracking problems in real-time. Moreover, due to the advances of hardware computing performance, [18, 4, 16, 30] solve textured object recognition problems based on the development of high-dimensional descriptors and optimized classifiers (e.g., randomized trees) [5, 10, 20, 26, 25], although there are view direction, illumination changes and rotation. Furthermore, [15] detects less-textured objects based on the dominant orientation templates [11], even in the case of saturation or shading within a captured color image. However, these computationally optimized approaches are constrained by the types of target objects, which would fail in cases where there are different constraints or error sources.

To overcome the above-mentioned problems—that is, that different error sources could produce failure—prior to the computationally optimized approaches, many methods originated and progressed by understanding and imitating how human eyes biologically process. In an early stage, [19] imitated on-cell and off-cell of receptive field of retina for edge detection. Since then, more recent approaches about object recognition and even edge detection still imitate the biological process of the human eye, such as the properties of receptive field [3, 13] and ganglion cells [32] of the retina. However, those biologically motivated algorithms also sometimes fail in cases that, for example, different objects have similar features. The reason is that human visual perception is affected not only by low-level visual information processing but also by other factors, such as experience and intention in a top-down manner.

To provide stable object recognition that is applicable in real situations and robust to the variations of target object type, we propose the unified visual perception model (UVPM) combining (bottom-up) low signal processing and (top-down) predictive information processing approaches. The proposed UVPM has academic plausibility by adopting the experimental results conducted from the field of cognitive, psychological science and brain research. Contrary to the studies of different research fields, we computationally implemented each component of the proposed model to prove its effectiveness. Moreover, based on the designed hierarchical LTM, the proposed model is hierarchically extendable from feature point matching to space-level recognition. Furthermore, the model is applicable to the AR scenario because the feature point matching itself can be applied to implement conventional AR methods. By retrieving hierarchically related candidates for a recognized object, the proposed model helps implement hierarchical context-aware wearable AR, as shown in Figure 1.

Figure 1: Concept diagram of context-aware wearable AR: The recognized object names are overlaid on the aligned positions of optical see-through HMD. The proposed model retrieves and gives more information about the recognized object or space based on the hierarchical memory structure.
The rest of this paper is organized as follows: Section 2 analyzes the previous models, utilizing contextual information for visual information processing. Section 3 gives an overview of the proposed model. The current status of the proposed model prototyping and a detailed description of the implementation are presented in Section 4. Section 5 discusses conclusions and future works.

2 RELATED WORKS

Human visual perception is affected by the context of situation; that is, relationships between objects and a space. For replicating the facts, there have been studies combining contexts with given visual information. For better object detection, [22] utilizes the probabilities representing the likelihood of an object that could be placed in a specific position of the image. However, if a scene image is changed or different types of objects exist within an unexpected scene, the use of the position probabilities might become ineffective. As the object position probabilities related to a captured scene help to control visual attention [22], the relationships among objects, as contextual information, increase the recognition accuracies by filtering out the unreasonable matches (e.g., among the initially recognized boat, sky and grass instances, contextual information replaces the boat with the airplane [9]).

The integration of context information and object recognition expands into total scene understanding. Based on the segmented and recognized object components, [17] additionally classifies the overall scene and then labels each component names on it. However, this could not be seen as a top-down approach even though those approaches utilize the contextual information: The so-called contexts are interpreted only from the given images. In this case, if the segments and filtered objects were unstably filtered out, then following other interpretations might possibly fail. To recover from those bad effects derived from the bottom-up approach, there have been some studies about the model for visual information processing or combining bottom-up and top-down approaches.

From the computer vision point of view, [7] models visual information processing in brain and insists that the right way for visual information processing is top-down (from coarse to fine) processing. There are some debates supporting the assertion from the field of cognitive informatics [31] and brain research [8]. [31] introduces a framework of human visual information processing based on a cognitive principles of human visual perception. Interesting results of [31] include that the brain carries out visual information processing in an abstract approach as a symbolic manner, and all visual information is represented and processed as visual semantic objects rather than direct images in LTM. In terms of brain research, [8] also proves the top-down facilitation of visual object recognition, triggered by the typical appearance for an object (described by symbolic form in [31]) and predictive information according to the relationships between an object and other objects around that object (described by semantic form in [31]).

Based on the previous cited works, we could ascertain that the visual processing should consist of bottom-up and top-down approaches for stable object recognition. In particular, in terms of a bottom-up approach, we need to consider how to configure the extracted features in a symbolic representation before focusing on making the features discriminable. Moreover, in the top-down approach, we need to consider how to make and manage the semantic form as predictive information. Based on the considerations, we propose a computationally unified visual perception model that conforms to proven cognitive science and brain research. Furthermore, we discuss the direction of improvement and its extensibility to context-aware wearable AR.

3 PROPOSED UNIFIED VISUAL PERCEPTION MODEL

As shown in Figure 2, the proposed UVPM is so designed that the WM and L/STM are mutually complementary for understanding visual information as a combination of top-down and bottom-up approaches. To be more specific, bottom-up approach is relevant to the WM, which is in charge of visual information processing, and top-down approach is relevant to the L/STM, which is in charge of predictive information processing. The WM of the model carries out a role of identifying the candidates of space and multiple objects which can be found in a space. For those identifications, the WM receives multiple sensory input data from various sensors such as light, compass, and location sensors (for space identification) as well as RGB-D imaging sensor (for multiple objects identification). Based on the multiple sensory inputs and a predefined hierarchical relation tree as a reference for the predictive information processing, the proposed UVPM processes the visual information.

In sequential order, at first, WM analyzes environmental context and identifies an initial candidate of a space from the various sensory data. Then, the model brings a part of the hierarchical relation of feature-object-space (F-O-S) tree from LTM to STM based on the identification, as shown in Figure 3. The entire F-O-S tree, stored in LTM, represents the hierarchical structure of the relations—among features, objects and spaces—that can be labeled as knowledge or experience, so that the use of inter/intraclass (F-O-S) relations increases recognition accuracies by refining the object recognition results taken from WM. For example, on a dark night, humans can identify an object if they know their location and that of other objects because they generally experience the situation and memorize the relationship, in the same manner of [7, 8, 31]. Thus, the F-O-S tree representing the knowledge of inter/intraclass relations allows the recognition tasks to overcome the ambiguity problem.

4 CURRENT STATUS OF UVPM IMPLEMENTATION

In this section, we describe the current status of prototyping the proposed model and a detailed description of the implementation. The adopted techniques or algorithms could be replaced with other algorithms providing more stable performance. At this stage, we made a prototype of the model to ensure that the proposed
concept of the model works well in a global aspect. For the implemented boxes of the prototype, we give detailed descriptions. For the not-yet-implemented boxes, on the contrary, we explain the meaning of the boxes in Figure 2.

**Environmental Context Analysis and Space Inference:** Humans can retrieve information related to the location where they are. However, in the manner of a computational model, location must be given by input sources. Thus, we assume that the multiple sensors (e.g., light, compass, location and gyro sensors) that are built into current mobile devices are made available to use. By analyzing the combination of the input data, we can infer a candidate of space, which is one of the preliminarily defined many spaces. The inferred space candidate is to be used as an input for bringing a part of the hierarchical F-O-S tree from LTM to STM because searching range of the whole tree needs to be reduced for efficient search.

**Raw Feature Detection & Video-based Object Clustering:** As described in both [19] and [23], first, raw features (called tokens in [19]) are extracted. Then, motion information helps group the extracted raw features as each different cluster having similar motions [23]. Based on the biological learning process of visual parsing described in [23]), the box of the proposed computational model detects saliency map as an imaged raw features. After that, it groups each salient region — region of interests (ROI) — by finding the closed contours rather than utilizing temporal information in this prototype of the model. The detected saliency map and initial ROIs are shown in Figure 4.

For the saliency map detection, we adopted the histogram-based contrast method (HC) described in [6]. The reason we chose the method among various state-of-the-art saliency map detection methods is that it requires low computational power while providing acceptable outputs. Based on the revealed implementation source code, we detected the saliency map, as shown in Figure 4(a). Then, we detected the candidates of the ROIs based on the contour detection method provided by OpenCV library [1]. If a detected contour was enclosed by other contours, it retrieved only the extreme outer contours for all the contours, as shown in Figure 4(b).

**Object-based Features Extraction and Object Recognition/Tracking:** As described in [8], the proposed model tried to imitate the concept of coarse representation, as a symbolic form, for the initial guesses of objects. For the concept, we adopted the bag-of-words (BoWs) method to represent an object. The reason we chose the BoWs is that the BoWs itself treats an image as a set of features [28] (meaning its generalized representation). For object-based features extraction, we use FAST corner detector [29] and SURF descriptors [4] for considering scale and rotation variations and real-time implementations. For discriminative and fast descriptor quantization, we use Randomized Trees as a codebook and a classifier [20, 27].

BoWs for a ROI can be formed, through the codebook, by using the extracted descriptors as inputs. After that, the BoWs are classified based on the pre-learned classifier. In this stage, classifier determines the recognized object by selecting the object label that maximizes the probability among others, as shown in Figure 5(a).

With all the probabilities for all the candidate ROIs, we initially filtered ROIs by selecting the regions where the probability is larger than the predetermined threshold. In advance of the consideration of the probability, we also filtered the ROIs by measuring the size and the number of feature points extracted within a ROI. By filtering out the ROIs, which has a small size of the region or small numbers of extracted feature points, we finally detected the filtered set of ROIs, as shown in Figure 5(b).

**Figure 5:** Examples of **Object Recognition/Tracking** as ROIs filtering: (a) simplified representation of the tree-based classifier (b) filtered ROIs based on size, number of features and recognition probability.

**From Long-term Memory to Short-term Memory:** As [7, 8, 31] insisted from computer vision, brain research and cognitive informatics points of view, top-down approaches for the predictive information are necessary to understand visual information as well as the low-level processing. Thus, we defined the relationships among objects based on an ontology that is relevant to LTM, as shown in Figure 6(a). The ontology is modeled based on Java-based Protégé [2]. Defined properties of each individuals are “nearby,” “on to,” “underneath,” “placed in” and “size.”

**Figure 6:** Examples of **Long-term Memory** and **Object Recognition/Tracking** as final recognition: (a) ontology as the representation of hierarchical relation and (b) final recognition results.

**Model-based Object Inference and Object Recognition / Tracking:** To confirm the effectiveness of LTM, we defined somehow strong assumptions for the relationships objects. To retrieve the information related to the initially recognized results from the knowledge-based framework, we used the TCP/IP socket because the WM of the model is implemented by different programming language (C++). In implementation of the prototype, we utilized only “nearby” property for the object recognition. Based on the ontology, we could filter out the ROI wrapping PC monitor (which is not registered in the object database) in Figure 5(b), as in Figure 6(b). As an initial prototyping result, we could confirm that the model stands a chance of providing more stable recognition results.
**Task Analysis:** This is a type of black box that manages and processes the human mind or memory. We assume that this certainly influences the development of LTM based on motivation, experience, knowledge and so on. However, at this moment, we have not yet found the apparent process of it.

**5 Conclusion and Future Work**

We proposed a computationally unified visual perception model, which combines bottom-up and top-down approaches. The model reflects the sequence of human visual processing based on the theories of cognitive informatics, brain research and psychological science. By implementing the prototype of each boxes of the proposed UVPM, we confirmed that the model stands a chance of providing stable recognition results and that there are development possibilities. Furthermore, it is still possible to say that this model is applicable to the AR scenario because it utilizes the feature point descriptors for making the symbolic form (BoWs) of an object. Thus, conventional feature-based approaches for matching points and 3D registration are possible based on the feature descriptors comprising the BoWs.

There is still much remaining work to proceed and complete the proposed model and its application to context-aware wearable AR. In particular, the direction of the algorithm flows and the links between the boxes could not be verified. Moreover, even though cognitive science or brain research is reflected in some boxes of the model, the implementation for other boxes did not go completely as planned. Thus, we want to complete the proposed model by strengthening the theory base from psychological and biological research areas and experimenting to prove its effectiveness. Moreover, with the optical see-through HMD, we plan to conduct an application of context-aware and AR based on the proposed model.

**Acknowledgements**

This work was supported by the Global Frontier R&D Program on Human-centered Interaction for Coexistence funded by the National Research Foundation of Korea grant funded by the Korean Government(MSIP) (NRF-2010-0029751).

**References**


