Object recognition in images by human vision and computer vision

Citation for published version (APA):

Document status and date:
Published: 01/01/2010

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.

Download date: 12. Jan. 2020
Object Recognition in Images by Human Vision and Computer Vision

Qunli Chen, Jan Dijkstra and Bauke de Vries
Eindhoven University of Technology
P.O. Box 513
5600 MB, Eindhoven
the Netherlands
Q.Chen@tue.nl, J.Dijkstra@tue.nl, B.d.Vries@tue.nl

Key words: object recognition, visual perception, illumination, grey contrast

Abstract: Object recognition plays a major role in human behaviour research in the built environment. Computer based object recognition techniques using images as input are challenging, but not an adequate representation of human vision. This paper reports on the differences in object shape recognition between human vision and computer vision. To this end, we explore an image-based recognition algorithm and a model-based recognition algorithm. Experiments were conducted using these two algorithms given images generated by Radiance lighting simulation software. The results reveal significant differences between human vision and computer vision given various conditions in the testing images and in the testing room.

1. INTRODUCTION

Human behaviour research is traditionally the field of environmental psychology ([1], [2], and [3]). From their perspective the built environment is a given condition. Architectural designers, however, design spaces to accommodate human activities. Recently, architectural and engineering researches develop models to evaluate building performance ([4], [5], [6], [7], [8], [9] and [10]). Human behaviour plays an important role in these
models. We need a dynamic model, which can simulate human behaviour using agent technology for building a performance evaluator. In this model, the visions and actions of the agent are determined by applying human visual perception simulation to mimic real physical perception. Therefore, our first step is to find an object recognition algorithm to perform the human visual perception task in the model.

Visual recognition of objects is one of the most challenging problems in computer vision. Object recognition in given images using computer vision is influenced by object shape and texture, environmental conditions and position of the object.

In studies on human vision, researchers often conduct their experiments using monitors. As a result, images are presented to the subjects. Compared with build a physical mock-up, this saves a lot of time for researchers. However, high dynamic range images captured by subjects are replaced by images in RGB colour space. This replacement means a loss of information. Until now, few researchers argued the validation of results obtained by using monitor-based experiment for human vision researches with RGB images.

In this paper, we first explore an image-based recognition algorithm and a model-based recognition algorithm. Then, experiments are conducted using these two algorithms on object shape recognition with objects under various conditions in given images. With the analyses of these experiments, conclusions are drawn on differences in object recognition by human vision and computer vision.

2. TECHNOLOGIES FOR OBJECT RECOGNITION

Within computer vision technologies, currently two main approaches can be applied, namely Image-based Object Recognition and Model-based Object Recognition. With given images, Image-based Object Recognition algorithms concentrate on interest points of the object, parts of the object, and appearance of the object whilst Model-based Object Recognition algorithms focus on edge detection and object shapes.

2.1 Image-based Recognition for Computer Vision

In the early days, object recognition could only be applied to some specific object with a large number of training examples. As new technologies were applied in this field, some advanced models were created which can be used to recognize objects of the same type. However these
models still need a large number of segmented training examples. Starting from the year 2000 new object recognition systems that can learn to recognize many objects (one at a time) were created. These systems are called constellation systems, which can recognize the informative parts as well as the spatial relationships of the objects ([14], [15]) with few training samples. Following the underlying idea of constellation models, Lowe created the Scale Invariant Feature Transform model (SIFT) ([16]). This model is considered a landmark in history of object recognition.

In recent years, some researchers turned back to look at object recognition problem from the biology science view, and obtained satisfying results. Among them, Serre developed a hierarchical system which can be used for the recognition of complex visual scenes, called Standard Feature Model of visual cortex (SFM) ([17], [18], [19], and [20]). This biological motivated object recognition system has been proven to be able to learn from few examples and provide good performance.

In the SFM model four layers are distinguished based on S (simple) computational units and C (complex) computational units. The function of the S unit is to combine the input stimuli with Gaussian-like tuning as to increase object selectivity and variance while the C unit aims to introduce invariance to scale and translation. These four layers are called $S_1$, $C_1$, $S_2$, and $C_2$. A brief description of the functions, input and output to the four layers are listed in Table 1.

<table>
<thead>
<tr>
<th>Layer Number</th>
<th>Brief Actions</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$ 1</td>
<td>Apply Gabor filters to the input grey-value image</td>
<td>Grey-value image</td>
<td>maps of various positions, scales, and orientations</td>
</tr>
<tr>
<td>$C_1$ 2</td>
<td>Use a max-like operation for each of the four orientations and each band</td>
<td>Bands of maps from $S_1$</td>
<td>Maximum for each bands ($C_1$ features)</td>
</tr>
<tr>
<td>$S_2$ 3</td>
<td>Patches generation (different scales in the same orientation)</td>
<td>Target images of $C_1$ format</td>
<td>$S_2$ patches</td>
</tr>
<tr>
<td>$C_2$ 4</td>
<td>Combine a max operation and the $S_2$ patches, find the scale and position invariant features</td>
<td>$S_2$ patches</td>
<td>Position &amp; scales invariant features ($C_2$ features)</td>
</tr>
</tbody>
</table>
Next some details will be given to illustrate the working principles underlying the SFM model. The first two layers of the SFM model are $S_1$ and $C_1$. These two layers are designed such that they correspond to the first cortical stages of V1 (the primary visual cortex) in the brain. In the $S_1$ layer, Gabor filters ([17], [19], and [20]) are used to analyze the input grey-value images. Gabor functions have been shown to provide a plausible model of the cortical simple cell receptive fields, therefore by making use of the Gabor filters the $C_1$ layer can simulate the neurons’ tuning property in V1 ([22]). To make sure all the filters fit with the property of the simple cells in V1 ([17], [19]), a careful selection is taken. After the selection, a final set of 16 filters at 4 orientations are left whilst the other filters are incompatible with the property of the simple cells in V1. Then these 16 filters are arranged to form a pyramidal form of scale, varying from 7 * 7 to 37 * 37 pixels in steps of two pixels. At the same time, four orientations of the Gabor filters (0°, 45°, 90° and 135°) are chosen to trace the number of $S_1$ units. Therefore there will be 64 maps (16 scales * 4 orientations) in 8 bands as the output of the $C_1$ layer (each band contains two adjacent filter sizes, so there are 8 bands of the 16 scales).

Researchers found that the complex cells in the cortex show tolerance to the position and the size ([23]): the receptive fields of complex cells are twice as large as simple cells, and complex cells response to every oriented bars or edges in their receptive fields. Based on these ideas, in $C_1$ layer, the position and scale tolerant $C_1$ units are built using a max-like pooling operation. Once again, parameters of the operation are set to meet complex cells’ tuning prosperities. The inputs to $C1$ layer are maps from $S_1$ layer, which are pooled together under the same orientation and scale band to increase the selectivity and invariance. For each band, the maps of each band member are sub-sampled with a grid. The size of the grid is determined by the index of corresponding band. For example, when band 1 is chosen, maps of the same orientation with filter size 7 * 7 and 9 * 9 are pooled together. According to the band of these two filters, a grid with cells of size 8 * 8 is used. For each cell of the grid, a local maximum is taken from the 64 elements. At the last stage the maximum value is chosen from the two scales of the same band.

In the next layer $S_2$, firstly a set of patches are generated. These patches contain target images of $C_1$ format with different sizes at random positions at all orientations. After pooling these patches together, in the learning stage, these patches are trained as the prototypes (features) of the $S_2$ units. Each of the $S_2$ units is used as the radial basis function (RBF) units during the object recognition process. Therefore with a given input, the $S_2$ unit response
depends in a Gaussian-like way on the Euclidean distance between this input and one stored prototype ([17], [19], and [20]).

The last layer, the $C_2$ layer deals with the shift and the scale again. This time, it aims to find the global max over all the scales and the positions for each $S_2$ patch. Effectively, a vector of maximums is generated by this procedure. The length of the vector is exactly the same as the patch number in $C_1$ layer. These maximums are position and scale invariant features (C$_2$ features). They are more general than the C$_1$ features, with which better recognition results are achieved. In the classification stage, the $C_1$ or $C_2$ response to a given image will be passed to a simple linear classifier for final analysis. For a more detailed explanation, see Serre’s paper ([17], [19], and [20]).

2.2 Model-based Recognition for Computer Vision

In comparison with Image-based recognition, Model-based Recognition represents the object data in a different mode. A survey of Model-based Recognition system is reported by Chin [24]. An important research field in Model-based Recognition is Geometric Modelling for Computer Vision, which focuses on 3D description (geometry, topology, colour and texture of the object, etc.) of the object in the model (e.g. CAD model). In Geometric Modelling for Computer Vision, the model description is compared with features extracted from the scene for recognition. In our research, we focus on Geometric Modelling for Computer Vision.

Various methods are applied in Geometric Modelling for Computer Vision [12]. Four commonly used methods are: Wireframe Models, Set-Theoretic Modelling, Boundary Representation and Desirable Model Properties for Vision.

In Wireframe Models, objects in the model are represented using edges and vertices. The edges and vertices of objects are stored in a list. Recognition in Wireframe Models is based on edge matching. This method always achieves good result when the presented models are solid models.

The other name of Set-Theoretic Modelling is computational solid geometry (CSG). In CSG, objects are assembled with primitive shapes (such as rectangle boxes, spheres, cylinders and cones) using set operators. There are three set operators, namely union, intersection and difference, which are similar to Boolean OR, Boolean AND, and Boolean NOT. Using CSG, a given model will be represented in a tree structure.
One other method in Geometric Modelling for Computer Vision is Boundary Representation. In comparison with CSG, which only uses primitive objects and Boolean operations for assembling objects, Boundary Representation applies more geometry. In Boundary Representation, an object is represented in a more complex structure, which contains information on each of the object’s faces, edges, vertices and how these elements are combined. Given an object in Boundary Representation, the description of this object is divided in two parts: topology and geometry. Topology shows the connectivity of the faces, edges and vertices of the object, while geometry gives the information on the exact shape and position of each of the edges, faces and vertices.

The last method in Geometric Modelling for Computer Vision is Desirable Model Properties for Vision. This method is applied when a solid model is used for vision purpose, for example, pairing the model features and the observed data features, or predicting the appearance of an object from any position. In the other words, this method involves techniques which reduce the information implied in the models (e.g. CAD models), such that the information found is similar to the information perceived by human vision. According to the human visual perception process introduced by Gibson [11], Sun introduced a vision algorithm using Desirable Model Properties for Vision method [13]. The artificial visual process used in Sun’s vision algorithm is similar as the process from the light rays to the object notion supported by the cognition rules. Given a 3D CAD model, a view can be rendered into a pixel array. Objects in the 3D space are rendered as the patterns in the pixel array. The object type information as well as its texture and material information is attached to each pixel in this array. Once an object is recognized based on its geometrical information in the pixel array, extra information of this object can be obtained simultaneously. In other words, in this algorithm, the object recognition task is narrowed down to object shape recognition.

2.3 Preliminary Conclusion

The Image-based approach and the Model-based approach are fundamentally different. The Image-based approach is appealing because it is very generic and does not require any specific information model. The Model-based approach can benefit from available geometric models, which probably leads to much more efficient computing. Most importantly in our research, we want to mimic human vision, and thus control the object recognition algorithm to obtain ideally the same result as in reality. In reality
human object recognition is affected by environmental conditions. Because Image-based recognition seems most promising, we set out for an experiment using the SFM technique explained before. In the experiment we compare computer object recognition using SFM with human object recognition under various environmental conditions.

3. EXPERIMENT DESIGN

Two series of experiments were conducted to research the influence of environmental conditions on human vision and computer vision.

In the first experiment, images were displayed on a LCD monitor on a desk in the middle of the testing room. Random levels were assigned to the room illumination (100, 200, and 300 lux) in the testing room, as well as the brightness (25%, 50%, and 75%) and contrast (25%, and 50%) of the monitor. Without guessing the object shapes, 31 subjects (14 males and 17 females) were asked to identify the shape of the objects in the images for each combination of environmental conditions.

In the second experiment, object recognition task was performed using SFM with given images. This time, images with triangle shapes were used as positive input whilst images with other shapes are negative samples to the SFM. Moreover, images of positive and negative input are divided in halves, half for training and the other half for testing.

Unlike existing experiments on object recognition, several environmental conditions are set in the Radiance model before rendering the testing images, namely grey level contrast and room illumination.

A total number of 150 grey scale images were generated with Radiance lighting simulation software under the same conditions (room dimensions, room illumination and grey level difference between the object and its background). Given a fixed grey value for the background, the grey level of the object in the image is increased from 0.01 to 0.05 (Radiance RGB value) in steps of 0.01. Distinct values of room illumination (18, 50 and 100 lux) were applied for each grey level in each image. Moreover, selective shapes (triangle, pentagon, irregular quadrangle, rectangle, square, and trapezium) are assigned to the object in the images.
3.1 Object Shape Recognition in images by Human Vision

In this experiment, human vision is under effect of many variables: room illumination and grey level contrast in the images, room illumination in the testing room, brightness and contrast of the LCD monitor, gender and age of the subjects, and whether or not the subject is wearing glasses. Data is collected from 31 subjects.

Report produced by correlation analysis showed that the correlation between the recognition and the variables gender, age, whether or not wearing glasses were not significant (p > 0.05), nor was the correlation between recognition and the room illumination in the testing room (p > 0.05). Accordingly, all attention was focused on the other variables: grey level contrast and room illumination in the images as well as the contrast and brightness of the LCD monitor.

Data collected in the experiments relate the subject’s performance to some physical aspects of the stimulus. A psychometric function can be fitted to the experimental data, such that the response threshold(s), which indicates the stimulus intensities required to produce a given level of performance, can be derived. A general form of the psychometric function is ([25], [26]):

\[ \Psi(x; \alpha, \beta, \gamma, \lambda) = \gamma + (1 - \gamma - \lambda)F(x; \alpha, \beta) \] (1)

Where \( x \) is the stimulus intensity, in this case \( x \) will be the variable grey level contrast, room illumination in the images, contrast or brightness of the LCD monitor. The guess rate \( \gamma \) is the subject’s response at zero stimulus intensity, and the miss rate \( \lambda \) describes how often the subject fails to detect a large stimulus intensity. Function \( F \) can often be derived as the cumulative distribution function of the underlying perceptual process that generates the data, such as the Weibull distribution [21]. We use the Weibull distribution:

\[ F(x; \alpha, \beta) = 1 - \exp \left(-\left(\frac{x}{\alpha}\right)^{\beta}\right), \quad x \geq 0 \] (2)

This distribution is often used to model phenomena that increase in probability with increasing stimulus intensity, and it is widely used to model discrimination and detection experiments.

In the experiment, some small factors, such as the various subjects’ eye heights when they were sitting in front of the monitor, and the time each subject used to get use to the testing environment, can influence their recognition performance. To reduce the effect of these factors, in the
experiment, each participant were given enough time on eye adaption to the environment, and the height of the seat was adjusted for each subject. Moreover, in the psychometric function, we assigned the guess rate $\gamma$ with a value 5% to introduce these factors. For the miss rate $\lambda$, we use the frequency that the subjects misrecognised the object shape when a corresponding variable was assigned with the highest level.

The response threshold is defined as follows:

$$\Psi_p^{-1} = \alpha^\beta \sqrt{\ln \frac{1-\gamma-\lambda}{1-p-\lambda}}$$

(3)

Note that $\gamma$ and $\lambda$ rise from guessing and lapsing, which are independent from the stimulus. As a result, the threshold value is usually defined as 50% of the Response distribution function (Response=0.5) that involves the desired percentage of subjects that notice the recognition. In our analysis, we also adopt this definition.

Constrained non-linear regression is firstly applied to obtain parameter values in the mentioned psychometric function for each variable. In the report only grey level contrast and room illumination have fairly $\rho$-square values. Based on the parameter values in the psychometric function for these two variables, our fitting curves are drawn in Figure 1 and Figure 2.

![Figure 1: Psychometric function for Grey level contrast](image-url)
Figure 1 shows the psychometric function of grey level contrast of the images presented. In the figure, grey level contrast has a very small guess rate $\gamma$ and miss rate $\lambda$, and it has a threshold of 0.012.

![Psychometric function for grey level contrast](image1.png)

*Figure 1: Psychometric function for grey level contrast*

Figure 2 shows the psychometric function of Room illumination intensity in the images presented. The curve in this figure indicates that, room illumination has fits and threshold of limited reliability because of insufficient data.

![Psychometric function for Room Illumination Intensity](image2.png)

*Figure 2: Psychometric function for Room Illumination Intensity*

3.2 **Object Shape Recognition in images by Computer Vision**

In experiments using computer vision for object recognition, testing images are the input to the SFM. We randomly choose triangle-shape object as the object we want the computer to recognize. Therefore, images with triangle-shape object are used as positive input and images with non-triangle-shape objects are the negative input. Both the positive input samples and the negative images sets are divided in halves, one for training and the other for testing.
This time, the conditions in the testing room have no effect on object recognition performance. The only variables that affect the recognition rate are room illumination in the image, and grey level contrast in the image.

In the first part of this experiment, images with all room illumination levels (18, 50 and 100 lux) and all grey level contrast levels (0.01, 0.02, 0.03, 0.04 and 0.05) are used as input. Table 2 shows the object recognition rate given different object shape as negative input. The average recognition rate in this table is around 45%.

<table>
<thead>
<tr>
<th>Negative Object</th>
<th>Triangle (positive)</th>
<th>Other shapes (negative)</th>
<th>Correct recognition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pentagon</td>
<td>8</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Irregular Quadrangle</td>
<td>8</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Rectangle</td>
<td>8</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Square</td>
<td>8</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Trapezium</td>
<td>8</td>
<td>7</td>
<td>9</td>
</tr>
</tbody>
</table>

Given the fact that two variables may influence the object recognition performance, separate experiments were conducted on finding their significance. In a previous experiment we learned that human recognition probability for room illumination and grey level contrast are around 27 lux and 0.005, therefore images with room illumination of 18 lux and grey level contrast of 0.01 are dropped from the input samples. In the second and third part of the experiment, we conducted our experiments with exclusion of these lower levels.

<table>
<thead>
<tr>
<th>Negative Object</th>
<th>Triangle (positive)</th>
<th>Other shapes (negative)</th>
<th>Correct recognition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pentagon</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Irregular Quadrangle</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Rectangle</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Square</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Trapezium</td>
<td>8</td>
<td>7</td>
<td>5</td>
</tr>
</tbody>
</table>
In the second part of the experiment, images with a room illumination value higher than 18 lux were used as input. From Table 3, the average recognition rates of triangle shape rise to 90% compared with data in Table 2. This indicates room illumination in the image is significant on object recognition by computer vision.

The third part of the experiment was carried out given all images with grey level contrast higher than 0.01. The result is shown in Table 4. In the table, data reveals grey level contrast between the object and its background is not as significant as room illumination in the image. Images with grey level contrast higher than 0.01 will give little improvement in recognition.

<table>
<thead>
<tr>
<th>Negative Object</th>
<th>Triangle (positive)</th>
<th>Other shapes (negative)</th>
<th>Correct recognition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
<td>Testing</td>
<td>Training</td>
</tr>
<tr>
<td>Pentagon</td>
<td>6</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Irregular</td>
<td>6</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Quadrangle</td>
<td>6</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Rectangle</td>
<td>6</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Square</td>
<td>6</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Trapezium</td>
<td>6</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

4. **SIGNIFICANCE**

With two series of experiments designed, a model-based recognition algorithm and an image-based recognition algorithm are applied to find the difference in object recognition in images rendered under various conditions using human vision and computer vision. The conclusions are drawn on the following aspects:

- Thresholds of grey level contrast in the image using LCD monitors for human vision research
- Effects of room illumination in testing images on object recognition using computer vision

5. **CONCLUSION**

In this paper, experiments were conducted to find the difference in human vision and computer vision on object recognition in images. Analysis
of the human vision experiment results show that it is feasible to design experiments on human visual perception with monitors given the precondition that proper value is set for grey level contrast in the images.

In further research, comparison of object recognition with human vision in two scenarios will be made. In the first scenario, room illumination and luminance on the monition in the testing room will be adjusted to meet the conditions in Radiance model. In the second scenario, random value will be given to the room illumination as well as the brightness and contrast of the monitor. From these experiments we expect to find more detailed information on the validity of computer-based vision experiments. We will report about that findings in the near future.

6. REFERENCES