Human and Machine Capabilities for Place Recognition: a Comparison Study

Carlos A. Martínez-Miwa, Mario Castelán, L. Abril Torres-Méndez and Alejandro Maldonado-Ramírez
Robotics and Advanced Manufacturing Group
CINVESTAV Campus Saltillo
Ramos Arizpe, Coahuila, México
Email: carlosmiwa_18@hotmail.com; {abril.torres;mario.castelan}@cinvestav.mx; alejandro.maldonado.ramirez@gmail.com

***Abstract*—This paper is aimed at exploring the way humans perceive previously visited places under challenging circumstances while driving a car. The term challenging refers to scenarios that do not contain specially rich visual information. We developed a series of experiments to investigate the performance of humans and computer vision algorithms, in order to recall locations in video sequences that were gathered with a camera mounted on a car. Our experiments reveal that the state of the art in machine vision fails against humans for identifying places where subtle changes occur, for instance, when illumination varies depending on the daytime. However, machines do present greater capabilities than humans when the reference image appears within a video sequence that does not exhibit spatial and daytime variations.**

Keywords*—*Place recognition; Human-machine capabilities; Computer Vision; Feature detection.

I. INTRODUCTION

Visual recognition of previous visited places is a fundamental part of our daily life. The study of how living beings recognize places, considering the fact that moving to one location to another is involved, has a long history in neuroscience [1][2]. Discoveries have provided a physiological grounding on the representation of spatial locations in our brain [3][4]. Our mind assembles "percepts" from memory (abstract thoughts) into internal images that are exactly experienced as images arising from the senses [5]. In other words, the mind feeds upon its own constructed images. We consider ourselves to be well adapted and skilled to navigate through even previously unvisited places if our interpretation about the environment matches a certain amount of already seen features. This might be because, we have mentally built an internal representation of what the action of navigating means in our mind, so that we associate past memories to the current navigation task by assigning the label: *"this is a new place"*. In this way, our brain could start the building of a visual representation for a new place. To this end, we may first need to select features that define such place in a unique way, though this selection may not result straightforward. We have been taught in our early years to pay conscious attention into what we were doing; when we do so, neurons cells start firing together and strong connections about recognizing a place occur. In addition, if perceptual changes in the environment exist due to conditions such as time of day, source of illumination, weather conditions, etc., the process of selecting good discriminative and invariant features that characterize that place turns complex. When we navigate a place for the first time, we seem more attentive on details that we believe will define it. We want those details to be distinctive enough so that a strong association is created. In that way, when we return to this location in the future, even when different conditions are present, the selected features that describe it can be fired up distinctively and thus achieve a precise recognition. Yet today, it is not clear how the learning process towards the detection of special features for place recognition is carried out in humans. However, by performing experiments we may be able to better understand human perceptual and cognitive abilities. Interestingly, from a psychophysics perspective there is not enough work that has addressed the problem of place recognition. An exception could be [6], where an approach to human perception-action, more appropriate for complex cognitive functions, such as object recognition and spatial cognition, has been studied through experiments in virtual envionments. The decision of using virtual reality is because of its ability to provide subjects with a level of sensory realism and dynamic sensory feedback that emulates their experiences in the real world. As far as computer algorithms are concerned, a survey has been recently published in [7], where an extense analysis about place recognition is presented. In this research work, we are interested in analyzing human performance for place recognition tasks compared to the performance of computational algorithms, which are based in combinations of detectors and descriptors. We are interested in evaluating how attention and previous knowledge is linked to perception when the goal is to recognize a place in challenging conditions. The combinations that were used in this work were chosen primarily due to their relevance in other computer vision tasks, such as appearance-based mapping [8], and their availability in the OpenCV Libraries. We have used features that are defined as real-valued vectors, such as SURF [9], SIFT [10] or KAZE [11]. SIFT and SURF have been widely used in the computer vision community as a benchmark for comparison using visual features, however KAZE has shown results that outperform both of them as shown in [11]. Other important branch of methods for extracting visual features are the ones defined in terms of binary values, which have shown a benefit, especially in terms of processing time. For example, ORB [12] and AKAZE [13] have been successfully applied to the problem of Real-time Vision-based Simultaneous Localization and Mapping (SLAM) as it is presented in [14]. BRISK [15] and ORB have also been applied to the task of efficient image retrieval [16].

Besides the above feature detectors and descriptors, we have also used algorithms that focus on detecting keypoints

 ISBN: 978-1-61208-609-5

without describing them. For example Good Features to Track (GFTT) [17], FAST [18] and STAR [19]. As we are comparing the capabilities of computers and humans for place recognition, we have included a method (AVA) [20], which is based in the influential work of Itti *et al.* [21] about computational visual attention. These kind of algorithms emphasize the detection of image regions that are likely to draw the attention of humans. The outline of this paper is as follows: in Section II, we evaluate human perception for recognizing already seen places while driving; Section III describes computer vision algorithms performance for detecting previously visited locations; Section IV presents a comparison between results of both, humans and machines; finally, conclusions derived form these experiments and future work are presented in Section V.

II. Experiments on Human Perception

Our experiment is aimed at exploring the way humans perceive previously visited places under challenging circumstances while driving a car. The term challenging refers to scenarios that do not contain specially rich visual information, for example, streets that do not present striking landmarks such as store facades, graffiti, advertisements etc. In this sense, we chose blocks that people could identify as belonging to a generic neighborhood of the city.

Three different video sequences were recorded, each at a different time of the day (7 h, 13 h and 19 h), from a 0.8 km route at a velocity of 30 km/h. The camera used was a GoPro Hero4, which was mounted on a Chevrolet Cruze vehicle. Figure 1 shows examples of images of the same place taken at different times of the day. For evaluating purposes, each of the recorded videos was cut in five 10 s fragments, and included on a game-like user interface designed on a Matlab GUI. Evaluation involved 60 participants: 30 women aged 34 ± 14 and 30 men aged 27 ± 6 (the average age of the total set of participants was of 30 ± 11). All of them were selected based on this profile: they had to be car drivers, be in a range of ages between 20 and 50 and their driver's license needed to be valid.

Each subject was given the following instructions: first, a 10 s video sequence of an urban environment was displayed. Every video was loaded only once, so that each participant had to focus her attention completely on that opportunity. As soon as the reproduction was over, a random (reference) image was shown to the subject. Then, she was asked to carefully watch that image and determine whether or not it represented a scene within the video. If the answer was negative, another reference image appeared. Conversely, if the user was sure that the place in the picture corresponded to a location exhibited in the video, she would have to search, one-by-one, for the video frame that best matched the reference image. This process was held three times per video, until a total of 15 video sequences were displayed. It is to note that every subject was given three trials before the experiment started, so that she could get familiar to both, the interface and the test. The procedure is described in Figure 2.

Figure 1. Images of the same place taken at different times of the day (Group 2). The top image was recorded at 13 h while the bottom image was recorded at 19 h.

Aware of the existence of cognitive bias, we decided to perform our experiments in a relaxing environment - a Mexican-style coffee shop, during off-peak times. As a way of motivation, every person was told that she would get a reward at the end of the test. We did not establish a control group as such as the only variable to measure was a yes/no reply. However, we carefully supervised that subjects were not randomly selecting answers. For example, at least 10 subjects did not correctly pick the two answers of the "obvious" Group 4. For these subjects, we also noted that they were easily distracted or eager to finish the experiment. For this reason we decided not to trust those cases, keeping only the results from the other 50 cases (25 men and 25 women).

A total of 45 reference images were presented to each subject. Four main sets were used: Group 1 contained 16 reference images that were extracted from the shown video, *i.e.*, the reference image appeared exactly as one of the frames in the video; Group 2 contained 12 reference images representing places in the video recorded at a different time of the day, where spatial, environmental or lightning variations occurred (see Figure 1); Group 3 showed locations that could be found in the complete route but that were not contained in the video (see Figure 3); the 2 remaining images (Group 4), were totally random, exhibited just to relax and reduce possible tiredness of the subjects.

Quantitative results obtained by this experiment are depicted through Figure 4. The color yellow represents a *Yes* answer and the red a *No* answer. The figure is divided into the four groups described above, where Group 1 appears at the top, Group at the middle and Groups 3 and 4 at the bottom of the figure. Note how, for Groups 1 and 2, it was expected that subjects identified the reference image within the shown video, thus a 100% success decision rate would have meant a full

 ISBN: 978-1-61208-609-5

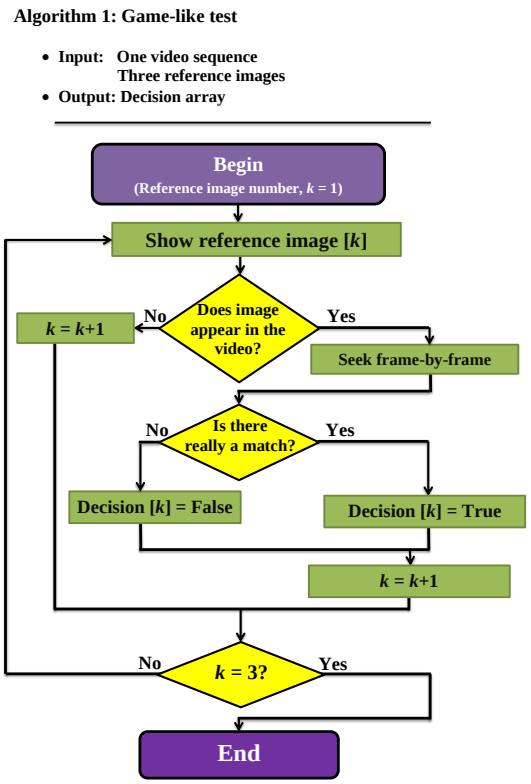

Figure 2. Flow diagram describing the steps followed by the participants. This represents one of fifteen cycles, for a total of 15 videos $\times$ 3 reference images = 45 attempts per subject. The average duration of each test was of 15 min.

(positive) yellow bar. As far as Groups 3 and 4 are concerned, the expected answers were negative, leading to a majorly red area. Percentages of success of 86 ± 7, 79 ± 11, 89 ± 6 and 99 ± 1 were achieved for Groups 1 though 4, respectively. The global mean percentage of success for the whole experiment was 86 ± 9 %.

TABLE I: FEATURES THAT, IN ACCORDANCE WITH PARTICIPANTS, HELPED THEM TO RECOGNIZE A PLACE. PV: PERCENTAGE WITH RESPECT TO THE TOTAL NUMBER OF VOTES. TS: TRAFFIC SIGNALS.

Feature	Buildings	Trees	TS	Other (20)	Color	Cars
PV(%)	25.97	22.07	16.88	16.23	15.58	3.24

As human attention is one of the main topics of our research, right after every participant completed the test, they were asked to choose from a list the features that grabbed the most of their attention. The list can be found with the results of this survey in Table I. It is important to note that the category "Other" contains all the elements that subjects expressed were helpful but were not listed in the survey. These items included benches, park games, wastelands, mountains, dumps, the driving way of the street, lampposts, etc.

III. Experiments on Machine Perception

For testing computational place recognition capabilities, the same 15 video sequences and 45 reference images as in the human perception evaluation were used. Each video was divided into 10 frames (1 frame per second), in order to follow the same way as human experiments. We tested 22 combinations of feature detectors and descriptors for a total of 9,900 comparisons. The methodology implemented, in all cases, is described as follows: first, a reference image and its 10 respective video frames were introduced as inputs to each algorithm. Then, the corresponding feature detector extracted the most relevant information from the reference image and each of the video frames. These features were explained by a descriptor, which assigned a unique feature vector for each image so that it could be identified among the rest. Next, the vector of the reference image was compared with the vector of each video frame, in order to find a best match. Once the 22 algorithms were evaluated with respect to all the 45 reference images, a threshold was applied to each detector-descriptor combination to determine wether this was a positive match or not. For threshold setting, we chose an 80% of the highest number of features found, *i.e.*, for each detector-descriptor combination there were a total of 450 evaluations. From these comparisons, we picked the image with the maximum number of correspondences as the 100% of success rate. In this way, we discarded as successful cases any images whose number of correspondences represented less than the 80% of the recorded maximum.

Figure 3. Example of an image that could be found in the complete 0.8 km route but that was not contained in the 10 s video (Group 3). The top image represents a scene in the video sequence, while the bottom one was the reference image shown to the participant.

The above process was evaluated for each of the 22 algorithms. The efectiveness of each pair of algorithms to correctly identify wether a reference image was included in its corresponding video sequence or not is shown in Table II. From the table it is noticeable how 7 out of the 22 combinations obtained the highest success rate while the poorest performance was exhibited by other 7 combinations. It is important to mention

 ISBN: 978-1-61208-609-5

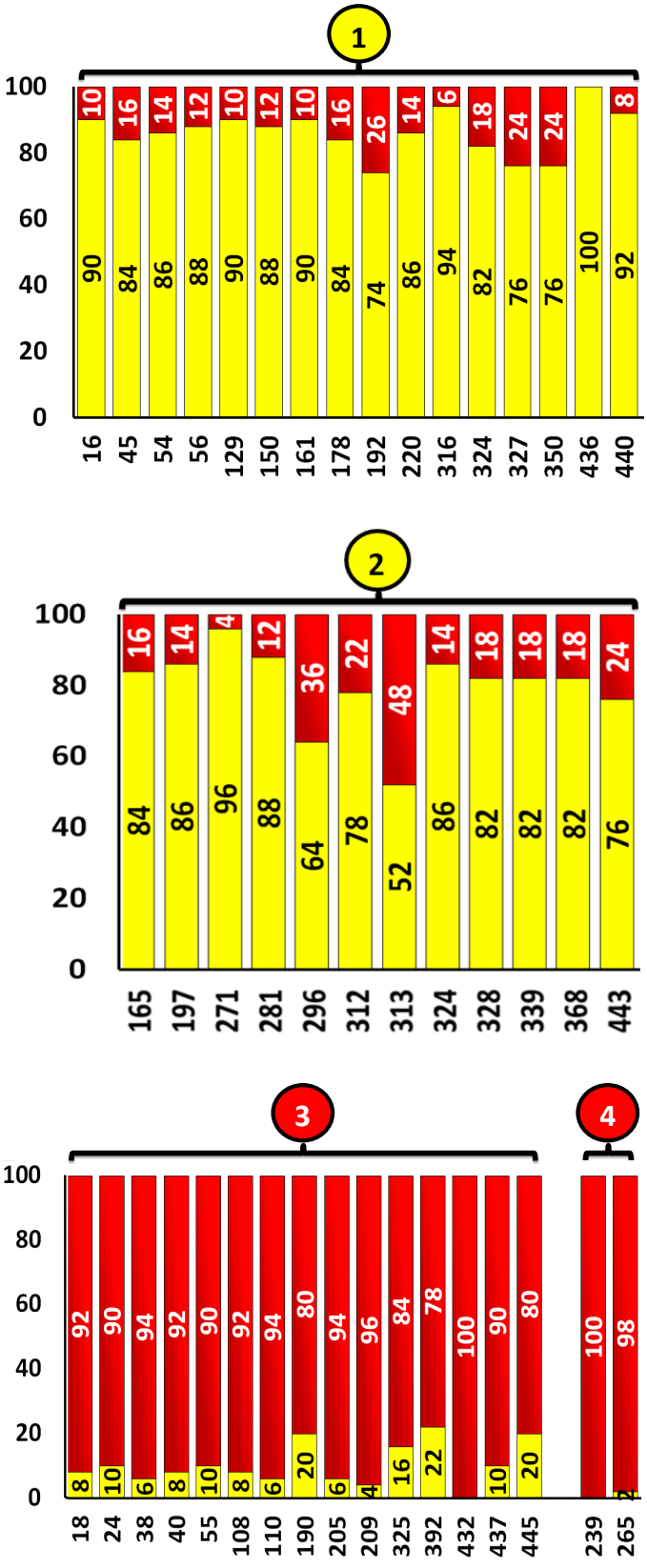

Figure 4. Human perception-Relationship between each reference image and the success rate of the experiment. For each diagram, the x-axis refers to the image number while the y-axis refers to the percentage of success rate.

TABLE II: LIST OF THE 22 DETECTOR-DESCRIPTOR COMBINATIONS USED IN OUR EXPERIMENT. THE GLOBAL PERCENTAGE OF SUCCESS WAS OF $54 \pm 41\%$.

Detector - Descriptor	Succes Rate (%)
AVA-ORB	73.33
AVA-SIFT	73.33
AVA-SURF	73.33
GFTT-BRISK	73.33
GFTT-ORB	73.33
GFTT-SIFT	73.33
GFTT-SURF	73.33
ORB-ORB	73.33
AVA-BRISK	51.11
STAR-BRISK	46.66
STAR-ORB	46.66
STAR-SIFT	46.66
STAR-SURF	46.66
KAZE-KAZE	44.44
AKAZE-AKAZE	42.22
BRISK-BRISK	40
FAST-BRISK	40
FAST-ORB	40
FAST-SIFT	40
FAST-SURF	40
SIFT-SIFT	40
SURF-SURF	40

that, for the 4 groups in the database none of the algorithms in the state of the art outperformed humans.

A further representation of the experiment is visually provided through Figure 5. The arrengement is similar to that of Figure 4, where results for Group 1 are shown at the top, for Group 2 at the middle, and for Groups 3 and 4 at the bottom of the figure. Here, it is to note that machines performed considerably worse than humans for Group 1, while for Group 2, where illumination changes happened, computer algorithms failed in all cases. Nonetheless, for Group 3, where humans obtained an 89% of success rate, machines achieved a complete 100%. This might suggest that the state of the art in computer vision is more effective at discriminating places that do not belong to a certain path than at recognizing previously visited places.

It is fair to comment further on the results presented in Figure 5, where machine capabilities for place recognition are depicted. For the cases of Groups 3 and 4, a 100% of correct answers for all detector-descriptor pairs was achieved; conversely, Group 2 denotes a complete fail among all of the combinations; thus, Group 1 is the only group that allows an individual evaluation of the performance for each pair. Besides, it is noticeable how the best 7 methods outperformed humans for Group 1 by achieving a 100% success rate. This means that the 7 following algorithms: AVA-ORB, AVA-SIFT, AVA-SURF, GFTT-BRISK, GFTT-ORB, GFTT-SIFT, GFTT-SURF and ORB-ORB, surpassed the capabilities of humans for Groups 1 and 3, although exhibiting a complete failure for Group 2.

IV. Discussion

From the experiments above, it is possible to discuss further on the capabilities of humans and machines for place recog-

 ISBN: 978-1-61208-609-5

nition in challenging driving environments. When comparing humans and machines, it is important to focus on certain nuances. For example, in Figures 4 and 5, it is noticeable how, while humans exhibit variations in their decisions, machines show uniform results. These discrepancies could be a consequence of human visual memory. Hayhoe [22] has defined human vision "as if our conscious experience were the ultimate end-product of visual processing". However, sometimes visual perception is not sufficient, and the need of a visual memory arises, which in accordance with Palmer [23] is "the preservation of visual information after the optical source of that information is no longer available to the visual system". All things in the world are separated by time or space so, as mentioned in [24], visual memory is needed to retain information about one thing in order to relate it to another thing. In addition, as things move around, they can occlude each other, and here is where visual memory helps to overcome the temporary loss of visual information. Nonetheless, although the process of recognizing a location by remembering and identifying just specific landmarks or features could produce a sufficient accurate and robust response [25], the lack of a reference might derive on a confussion for humans.

For the case of machines, images and videos are processed at a uniform resolution, so there is no notion of saccades to shift the center of high-resolution processing. Thus, there is no need for visual memory as a buffer for integration of information over saccades [24]. The behavior computer vision methods demonstrated appears to be similar between them when the task resembles image retrieval. However, as computer vision methods mostly rely on data, when there exists subtle changes between two images, machines struggle to find correspondences, even if both images are taken from the same location.

Outstanding cases of human perception can be analyzed in Figures 6 and 7. Machine perception cases are not shown due to the uniformity of its results, *i.e.*, for each group all methods achieved either 100% accuracy or 100% failure.

The best and worst performances for human perception in Group 1 are represented by Figure 6. At the first row it can be noticed how the existence of few, but representative items in the environment, such as a white fence, a pink wall and a tree with a particular shape (which can be associated with common day things like a corn) may help humans to remember and identify previously visited places. However, the second row of the figure suggests that if there exist similar features in two different places, or too many elements within a scene, subjects are prone to get confused. For example, in this case, the presence of trees, or a similar color of fences and walls in both images, might had been the reason of failure among participants.

Similarly, Figure 7 describes the best and worst results for humans in Groups 2 and 3, respectively. The top row corresponds to the best performance. Again, it can be observed that if there exist just a few characteristic elements in a scene, it could be simpler for humans to remember and recognize such scene, even if daylight or spatial variations occur. In the

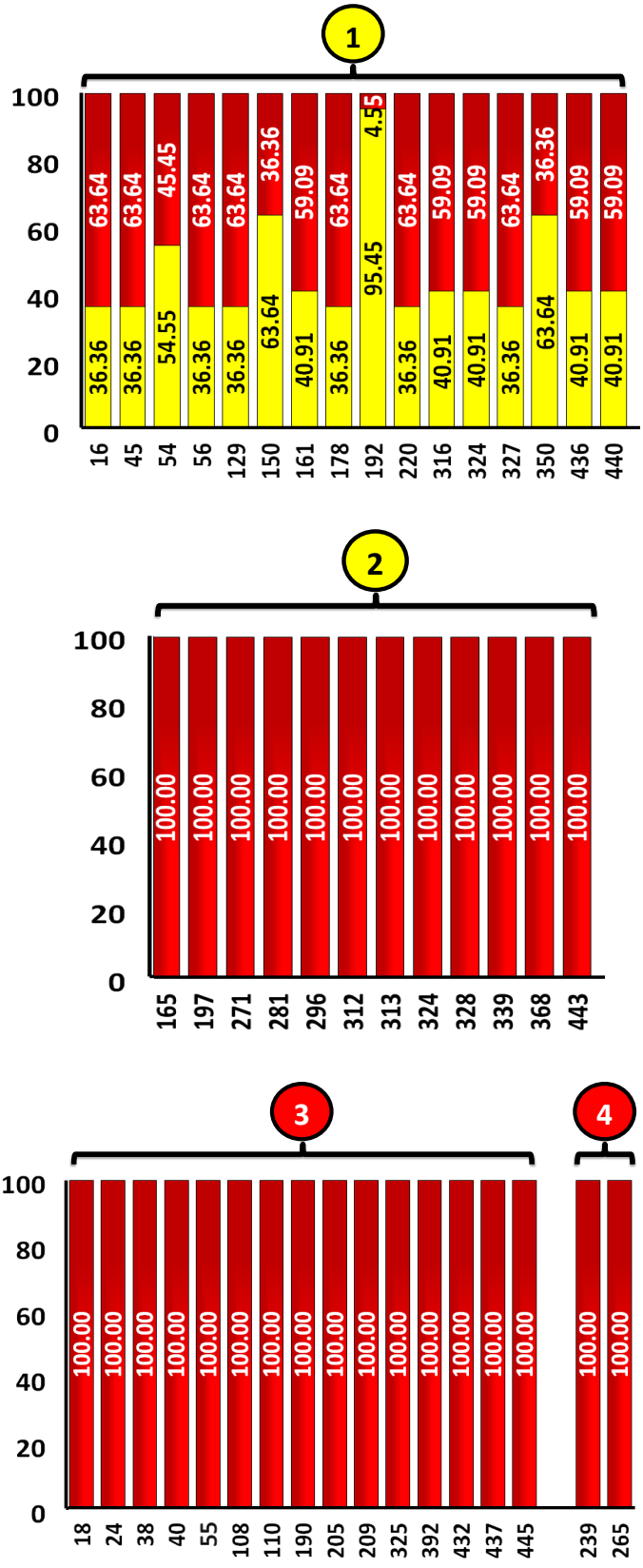

Figure 5. Machine perception-Relationship between each reference image and the success rate of the experiment. The *x*-axis represents the image numbers, while *y*-axis the percentage of success rate.

Figure 6. Best and worst cases for humans in Group 1 (top and bottom row respectively).

Figure 7. Images that represented the best performance for humans in Group 2 (top row), and the worst in Group 3 (bottom row).

figure, a single house at the left of the road, and the absence of items at the right, appeared to be enough for subjects to identify the location. Nevertheless, the bottom row shows, as it was mentioned above, that the presence of similar features can affect human perception. Here, we can see that the existence of fences that look alike could had been a cause for subjects to fail.

V. Conclusions and Future work

This paper is aimed at exploring and analyzing the way humans perceive previously visited places, while driving through urban environments under challenging conditions. An experimental setup was designed to evaluate human capabilities for place recognition. The same experiments were tested on state of the art visual place recognition algorithms. From experimental results, it can be observed that humans demonstrated a greater ability to identify scenarios with subtle changes, such as illumination or spatial variations, as opposed to machines, which did not accomplish positive results at all for these cases. However, machines outperformed humans when the problem became that of image retrieval, *i.e.*, when the reference image appeared without changes within the video.

As a future work it would be interesting to study in depth the weaknesses of computer vision algorithms so as to improve their robustness from the way humans perceive subtle changes.

References

[1] M. Riddoch and G. Humphreys, *Neuropsychology of visual perception*. Hillsdale: Lawrence Erlbaum Associates, 1989.

[2] R. G. Golledge, *Do people understand spatial concepts: The case of first-order primitives*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1992, pp. 1–21.

[3] J. O'Keefe and L. Nadel, *The Hippocampus as a Cognitive Map*. Oxford: Clarendon Press, 1978. [Online]. Available: http://hdl.handle.net/10150/620894

[4] A. D. Redish and D. S. Touretzky, "Cognitive maps beyond the hippocampus," pp. 15–35, 1997.

[5] J. F. Sowa, *Conceptual Structures – Information processing in mind and machine*. Addison-Wesley Systems Programming Series Reading, 1984.

[6] H. H. Bulthoff and H. A. van Vee, "Vision and action in virtual environments: Modern psychophysics in spatial cognition research," in *Vision and Attention*. Springer, 2001, pp. 233–252.

[7] S. Lowry *et al.*, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2016.

[8] M. Cummins and P. Newman, "Fab-map: Probabilistic localization and mapping in the space of appearance," *The International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.

[9] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[11] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, "Kaze features," in *European Conference on Computer Vision*. Springer, 2012, pp. 214–227.

[12] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *Computer Vision (ICCV), 2011 IEEE international conference on*. IEEE, 2011, pp. 2564–2571.

[13] P. F. Alcantarilla and T. Solutions, "Fast explicit diffusion for accelerated features in nonlinear scale spaces," *IEEE Trans. Patt. Anal. Mach. Intell*, vol. 34, no. 7, pp. 1281–1298, 2011.

[14] L. Caramazana, R. Arroyo, and L. M. Bergasa, "Visual odometry correction based on loop closure detection," in *Open Conference on Future Trends in Robotics (RoboCity16)*, 2016, pp. 97–104.

[15] S. Leutenegger, M. Chli, and R. Y. Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2548–2555.

[16] Y. Uchida, S. Sakazawa, and S. Satoh, "Image retrieval with fisher vectors of binary features," *ITE Transactions on Media Technology and Applications*, vol. 4, no. 4, pp. 326–336, 2016.

[17] J. Shi *et al.*, "Good features to track," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*. IEEE, 1994, pp. 593–600.

[18] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," *Computer Vision–ECCV 2006*, pp. 430–443, 2006.

[19] M. Agrawal, K. Konolige, and M. R. Blas, "Censure: Center surround extremas for realtime feature detection and matching," in *European Conference on Computer Vision*. Springer, 2008, pp. 102–115.

[20] A. Maldonado-Ramírez and L. A. Torres-Méndez, "Robotic visual tracking of relevant cues in underwater environments with poor visibility conditions," *Journal of Sensors*, vol. 2016, 2016, 16 pages.

[21] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[22] M. Hayhoe, "Visual memory in motor planning and action," in *The visual world in memory*. Psychology Press, 2017, pp. 117–139.

[23] S. Palmer, *Vision Science: Photons to Phenomenology*. MIT Press, 1999. [Online]. Available: https://books.google.com.mx/books?id=mNrxCwAAQBAJ

[24] S. Lallee, C. Tan, and B. Mandal, "Vision and memory: Looking beyond immediate visual perception," in *Computational and Cognitive Neuroscience of Vision, Cognitive Science and Technology*. Springer, 2017, pp. 195–219.

[25] C. Siagian and L. Itti, "Impact of neuroscience in robotic vision localization and navigation," in *Computational and Cognitive Neuroscience of Vision, Cognitive Science and Technology*. Springer, 2017, pp. 235–276.

 ISBN: 978-1-61208-609-5