Biologically Inspired Vision and Touch Sensing to Optimize 3D Object Representation and Recognition

Ghazal Rouhafzay, Ana-Maria Cretu, and Pierre Payeur

3D representation and recognition of objects are two pivotal steps for autonomous robots to safely explore and interact with an unknown environment and manipulate objects. 3D modeling can be beneficial in different robotic applications such as object grasping, pose estimation, robot navigation and localization. Real-time data acquisition and accurate object representation are essential in the context of such practical applications. On the other hand, the recognition of the objects in an environment is indispensable for situational awareness and for enabling the robot to interact effectively with complex environments.

Robot vision can be considered as the most informative and reliable sensing modality in autonomous robots. Nevertheless, vision fails to work properly in a number of situations including low light environments, cases where an object is occluded or is out of the camera's field of view, and situations in which objects are not visually distinguishable. Tactile sensing, as an indispensable element of dexterous robotic manipulation, can be efficiently integrated with other sensory modalities, in particular with vision, to increase the reliability of an autonomous robot. It makes available a wide range of information on objects including surface properties such as roughness, texture, vibration, temperature, local shape, etc., all important features that can contribute to better identifying an object. Moreover, a combined use of vision and touch in humans was demonstrated to facilitate manipulation, grasping and handling of objects, and could therefore be exploited to increase the efficiency of autonomous robots in a variety of tasks. However, visuo-tactile integration and the creation of efficient computation methods to help a robot successfully recognize and manipulate the objects it is interacting with remains a challenging issue.

A huge research effort has been invested in the literature to efficiently integrate the two sensing modalities. Nevertheless, all currently published works tackling visuo-haptic interaction only use visual data to increase the spatial resolution of tactile data, to resolve conflict situations, such as cases where the tactile information is faulty, or conjunctly use tactile and visual data to recognize objects. Considering the fact that the acquisition and processing of tactile data itself is a time-consuming task, such approaches for visuo-tactile integration are associated with a high computational cost, thus making them very difficult, if not impossible, to use in real-time interaction scenarios. Alternatively, the sophisticated cognitive skills of the human brain and its patterns of natural intelligence have encouraged scientists to develop biologically inspired computation techniques, bringing automatic processing capabilities to computers and robots.

Referring to biological research, we can draw three main conclusions about the interaction and collaboration of visual and haptic sensory modalities: 1) Tactile salient features also attract visual attention to their location [1]; 2) a combined use of vision and touch works more efficiently compared to cases where vision and touch are exploited separately [2]; and 3) visual and tactile object recognition rely on similar processes in terms of categorization, recognition and representation [2]. These conclusions suggest that visuo-tactile integration is a promising solution to optimize the process of object modelling. Moreover, visual data, in the form of salient regions acquired by a model of visual attention (according to 1) can be employed to guide the process of tactile data acquisition. Furthermore, visuo-tactile integration can be performed (according to 2 and 3) at a higher (perception) level based on similarities between the two sensing modalities. Since collecting large datasets of tactile data for training a model is a much more complex task compared to visual data, it is expected that a transfer of learning from vision to touch can both enhance the performance of tactile object recognition and amalgamate visual and tactile data processing units in robots. This paper presents research initiatives performed by the authors to validate these biologically inspired assumptions and efficiently merge measurements from different instrumentation technologies in a framework to operate in the context of practical robotic tasks that involve 3D object representation and recognition.

Techniques for Visual and Tactile Data Acquisition

As illustrated in Fig. 1, the framework makes use of visual and tactile data acquired over the surface of objects. A variety

1094-6969/21/$25.00©2021IEEE

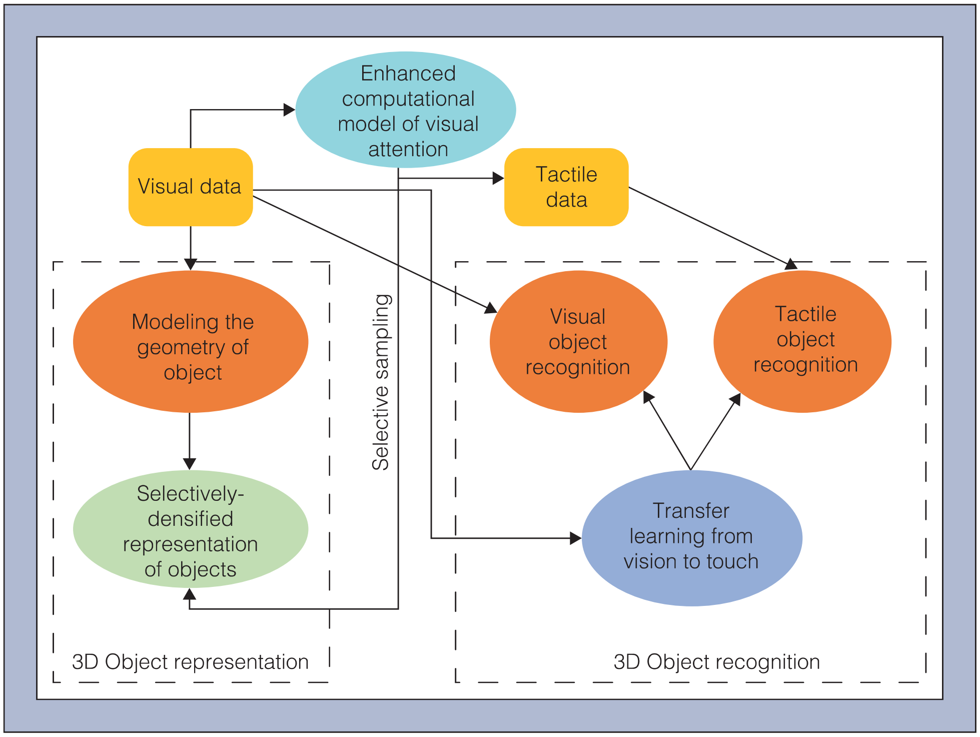
Enhanced computational model of visual attention
Visual data
Tactile data
Modeling the geometry of object
Selectively-densified representation of objects
Selective sampling
Visual object recognition
Tactile object recognition
Transfer learning from vision to touch
3D Object representation
3D Object recognition

Fig. 1. Overall framework for combined use of vision and touch in 3D object representation and recognition.

of instrumentation techniques are nowadays available for acquisition of the geometry of 3D objects; most of them rely on either time-of-flight sensor [3] or triangulation-based systems for modelling. Radio detection and ranging, Light Detection and Ranging (LiDAR), shaped light pulse, and sound navigation and ranging (sonar) are different types of time-of-flight sensors transmitting and receiving back narrow beam signals in order to compute the distance between sensor and object points. They can be used to acquire data on 3D objects through sweeping the beam over the object surface. Triangulation based systems basically take advantage of different viewpoints of objects from which specific features are sensed and matched. Stereo vision, structured light systems and spacetime stereo are examples of triangulation-based systems for 3D geometry acquisition. Stereovision is a technique where two images from different viewpoints are acquired from an object, which are then used to determine correspondences between the images [4]. Since no energy beam is emitted toward the object, stereo vision is referred to as a passive modeling solution. In structured-light systems, a light pattern is projected onto the object surface from one viewpoint. An image taken from a different viewpoint can be used to match corresponding points and compute the 3D coordinate of surface points [5]. Spacetime stereo systems project an arbitrarily varying pattern on the surface of object. Subsequently, they perform feature matching using two video streams, taken from different viewpoints over specific space-time windows. Such modeling systems are more appropriate for moving or deformable objects [6].

In the case of tactile data acquisition, existing tactile sensors characterizing surface deformations and texture of objects can be categorized into seven main types according to their transduction mechanism namely; resistive, capacitive, piezoelectric, optical and Organic Field Effect Transistor-based (OFET) sensors [7], acoustic, inductive and magnetic [8]. They produce a 2D image describing the surface of touched objects. Other technologies of tactile sensing relate the sensors characterizing the position, acceleration, vibration or force-torque sensors that sense the amplitude and direction of an exerted force. Among the existing technologies, resistive tactile sensors or more specifically Force Sensing Resistor arrays (FSR) are widely used for object recognition and have been selected in this research [9].

An Enhanced Computational Model of Visual Attention

At the heart of the proposed framework for combined use of vision and touch for 3D object representation and recognition (Fig. 1) is a computational model of visual attention. When looking at daily encountered visual scenes, the human visual system instantaneously processes the huge amount of available perceptual information in order to select a subset of relevant and required stimuli. This procedure of selection or inhibition of perceptual information is referred to as visual attention. Computational models of visual attention, which attempt to mimic the behavior of the human visual attention system, can be categorized into bottom-up and top-down models. Top-down attention mechanisms obey different types of cognitive factors in scene exploration such as, expectations, or searching for a specific target and prior knowledge, while bottom-up models rely on a set of effective features under free viewing conditions. For the latter, several researches from the field of neuroscience and psychology have identified contributing features in deployment of visual attention such as color opponency, contrast, curvature, edge, entropy, intensity, orientation and symmetry. As such, classical computational models of visual attention [10] build upon these features and apply the center-surround operations exhibited in human receptive field onto them to produce so-called conspicuity maps. Fig. 2 illustrates an example of conspicuity maps created based on a center-surround operation on a list of effective features for the guidance of vision as obtained by an enhanced model of visual

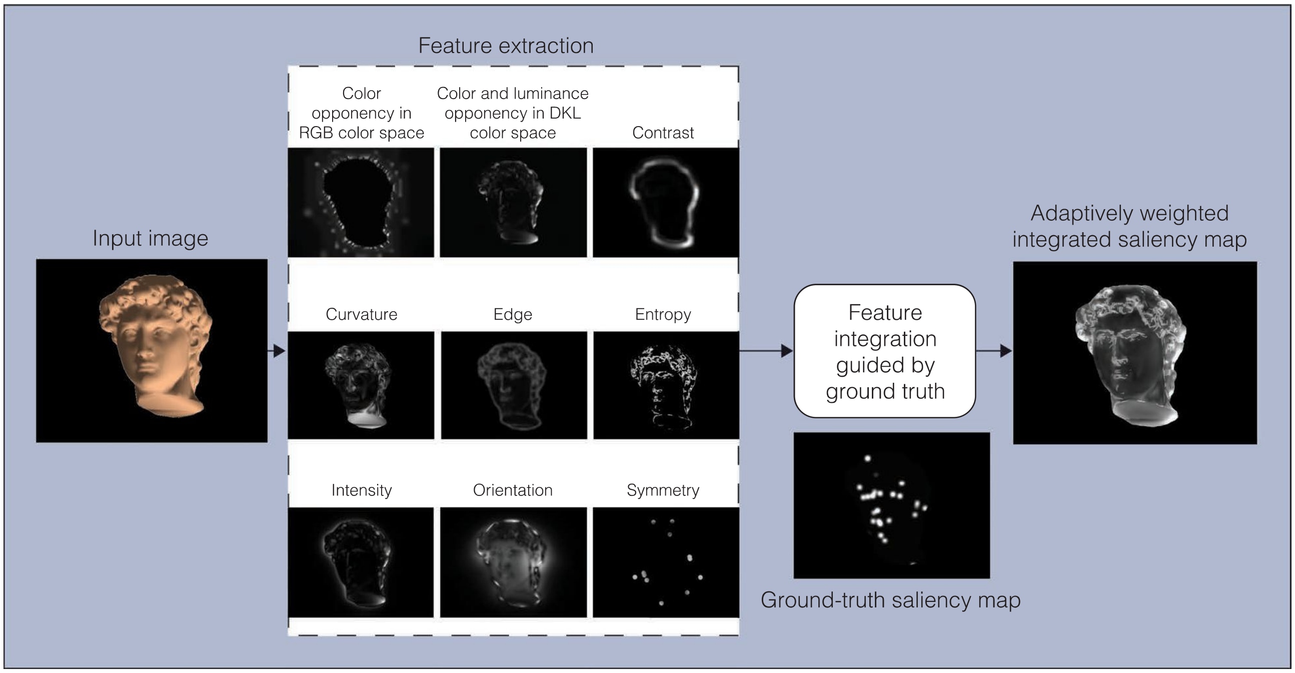
Feature extraction
Color opponency in RGB color space
Color and luminance opponency in DKL color space
Contrast
Input image
Curvature
Edge
Entropy
Intensity
Orientation
Symmetry
Feature integration guided by ground truth
Adaptively weighted integrated saliency map
Ground-truth saliency map

Fig. 2. Enhanced model of visual attention.

attention proposed in [11]. In this figure brighter pixels correspond to regions where higher attentional resources should be allocated.

Multi-feature integration is the next phase in the computation of a saliency map, which produces the final output of the computational visual attention model, highlighting saliencies as bright regions on a black background in 2D. In the literature, the conspicuity maps are generally integrated by averaging [10]. However, in an attempt to better guide attention, feedback from users (in form of a list of salient vertices identified over the surface of 3D objects) can help experimentally determine the contribution weight of each feature in integration process. This allows features that are more conform to human input to contribute more to the final saliency map. As such, for our enhanced visual attention model, a ground-truth saliency map is generated by casting a round Gaussian area (inspired by anatomy of receptive field in human vision system) around the salient vertices identified by users. Subsequently, a similarity metric between the ground-truth saliency map and each feature map is computed to produce a weighting scheme. Fig. 2 demonstrates an example of ground-truth saliency map and the final adaptively weighted saliency map. Experimental results identified curvature and entropy as the most prominent features in guidance of visual attention for a set of mono-colored objects rendered against a black background [11].

Another possible approach for feature integration is to adopt a machine learning technique to learn the location of salient points based on extracted features [11]. In such an approach, the process of saliency prediction is formulated in form of a binary classification problem. Subsampled versions of conspicuity maps are fed as input attributes to a Support Vector Machine (SVM) to predict if the associated pixel represents a salient region or not. A prediction accuracy of 86.08% is achieved.

Optimized 3D Object Representation

The fast rendering capability of triangular meshes has made them one of the most popular techniques for 3D object representation in robotics, virtual environments and computer graphics. Nevertheless, when the geometry of the object is complex, an excessive number of triangles is called for to achieve an accurate representation. The real-time creation and maintenance of an object scene containing several objects with enormous number of triangles becomes quickly impossible in robotic applications due to on-board memory limitations. Moreover, despite the fact that novel robotic platforms tend to exploit graphics processing units (GPU) to achieve real-time computation, the current technology of portable GPUs can overheat rapidly, calling for optimization of computational costs. This explains the interest in creating compact object representations, that can be efficiently stored and used, but that are also accurate, particularly in the areas that define the most predominant geometrical properties of the objects, and thus are more useful in object recognition and manipulation tasks. Computational models of visual attention can be advantageously engaged in this process to determine salient regions of object meshes which are to be represented at a higher resolution. These salient regions correspond to the most evident geometrical properties of the object (Fig. 3). For this purpose, the entire surface of a 3D object is first scanned in form of images captured from different viewpoints. The previously discussed model of visual attention is then computed over each acquired image to produce a saliency map. The next step is to derive salient points from salient regions. Since an over-concentration of salient points within a region will degrade

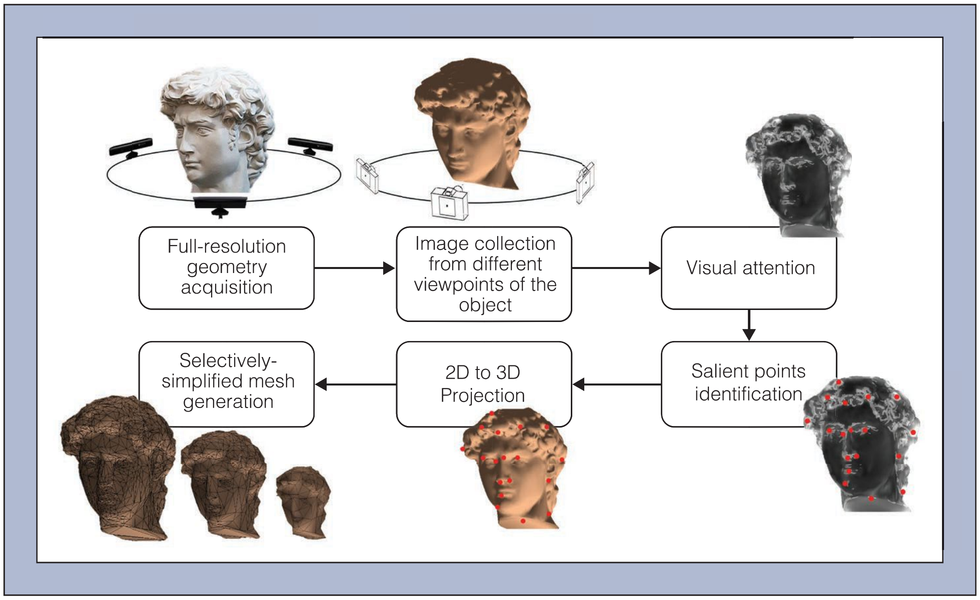
Full-resolution geometry acquisition
Image collection from different viewpoints of the object
Visual attention
Salient points identification
2D to 3D Projection
Selectively-simplified mesh generation

Fig. 3. Selectively simplified 3D modeling using the enhanced model of visual attention.

the quality of a simplified mesh, a non-maximum suppression scheme is applied to avoid a large number of salient points occurring within the same neighborhood over the mesh of the object. Salient points are finally projected to 3D as a list of vertices which should be preserved while other vertices remain accessible to Qslim algorithm for simplification [11]. Qslim algorithm is one of the most popular and efficient simplification algorithms where vertex pairs are iteratively contracted while maintaining a surface error metric. It simplifies 3D objects uniformly without considering the minor local features which are decisive in the representation of an object, justifying why the important salient areas are not subject to simplification. Using the proposed approach, different Level of Details (LoD) representations of objects can be generated based on the distance between a robot and the object (i.e., closer objects at higher resolution). Our experiments conducted over a dataset of 43 objects suggest that an average compression rate of 95.74 % can be achieved without distorting the object.

3D Object Recognition

3D object detection and recognition using robot-mounted cameras is a relatively well-established task in robotics. Literature on 3D object recognition can be broadly grouped into conventional methods that rely on feature extraction from segmented images and use a classifier to recognize objects, and deep learning-based method where feature extraction, object localization and recognition are all performed using a deep convolutional neural network (CNN) based architecture. Conversely, tactile object recognition can be performed either through dynamic or static touch. While dynamic touch relies on data acquisition through moving the sensing probe over the surface of object, a static touch is performed with an immobile tactile sensor contacting an object. Static touches capture local shape of an object, including small scale deformations on its surface, as illustrated in Fig. 4. The acquisition of static tactile data requiring the movement

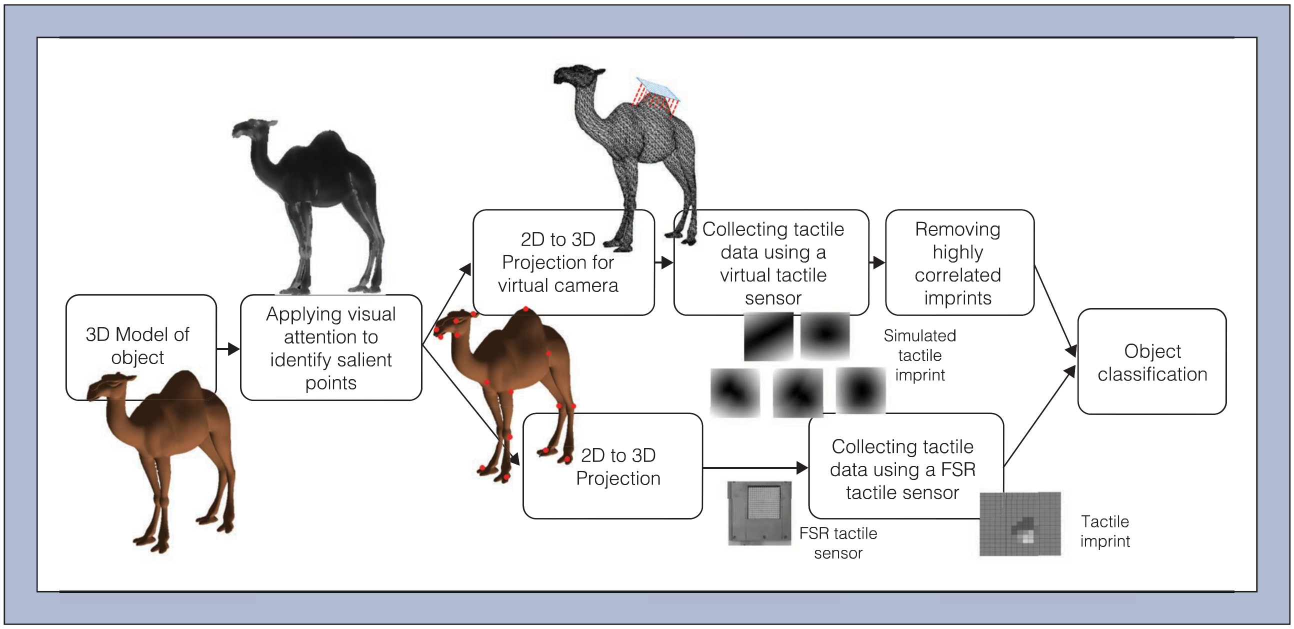
3D Model of object
Applying visual attention to identify salient points
2D to 3D Projection for virtual camera
Collecting tactile data using a virtual tactile sensor
Removing highly correlated imprints
Simulated tactile imprint
Object classification
2D to 3D Projection
Collecting tactile data using a FSR tactile sensor
FSR tactile sensor
Tactile imprint

Fig. 4. Object recognition from haptic glance.

and positioning of the tactile sensor and then executing a direct contact with the object is a tedious task. In terms of tactile object recognition, machine learning solutions (e.g., neural networks, self-organizing maps, kNN classifiers) tend to be solutions of preference, regardless of the source of acquired tactile data, the features extracted, and of the type of sensor used. As mentioned in the previous sections, in our work we used a Force Sensing Resistor (FSR) array to acquire tactile data and employ machine learning solution to classify selectively acquired tactile data, as detailed in the next section.

Visually Guided Selective Tactile Data Acquisition for Object Recognition

Inspired by the human visuo-haptic integration principle, in our proposed framework, we examined the use of the visual attention model to identify a series of interest points to determine the location where to collect tactile data over the surface of a 3D object to allow for the recognition of the probed object based on a limited set of such tactile imprints (i.e., from haptic glance). The value of this approach is that it avoids the tedious complete tactile acquisition process by identifying only a limited number of probing points from which tactile data is collected. Psychological studies about haptic perception suggest that humans are able to recognize objects among a small set promptly by a brief haptic exposure to a limited number of local tactile cues (i.e., from a “haptic glance” [12]). In this context, haptic glance can be defined as a short and static contact between the fingertip and the object of interest. The acquired tactile information by haptic glance is a combination of kinaesthetic signals sensed by joints, tendons and muscles as well as cutaneous cues detected by human skin. While kinaesthetic signals basically determine the location and position of the fingertip when touching the object, cutaneous cues provide information about local deformations and texture properties.

Experiments are conducted both on tactile imprints from a physical 16 × 16 FSR sensor and on simulated imprints using a virtual sensor [9], based on the working principle of a real FSR, on physical objects for which accurate 3D models were available. Since collection of tactile imprints is a tedious task in nature, the simulated tactile sensor was used for a proof of concept prior to experiments with a real sensor. When an external force is applied to an FSR sensor while the sensor is in direct contact with the surface of the object, the elastic layer of the sensor captures the geometric profile of the object. A 2D deformation profile, which forms the tactile imprint, is subsequently produced by the transducers in the sensor, as shown in Figure 4. Similarly, the virtual tactile sensor captures the geometry profile of the object surface, by measuring the distances between the elements on the sensor and the surface of the virtual object when the center of the sensor is in contact with the object [9].

Once tactile data are collected a set of features are extracted from them before classification. Feature extraction by wavelet decomposition resulted in a good performance on tactile images acquired in form of a 2D deformation profile by a tactile sensor [13]. Among the conventional classifiers from the literature, k-Nearest Neighbor (kNN) and Support Vector Machines (SVM) were demonstrated to be among the best options for tactile object recognition from haptic glance which usually relies on small datasets for the training phase [13]. Classification accuracies obtained by kNN and SVM are reported in Table 1. Experiments conducted using a real tactile sensor confirm the success of visual attention in optimizing the process of tactile object recognition, with classification accuracies on average 22.9% lower than those obtained through simulation for kNN and SVM [13].

Engaging kinaesthetic data in form of the probing location can improve the recognition rate up to 14.86% [13]. A comparison of experimental results suggests that for tactile datasets of equal size, the recognition rate when tactile data are collected from visually salient locations are up to 22.66% higher than the case where tactile data are acquired by a blind touch (random determination of tactile probing location) [13]. This supports the fact that visual attention allows selection of locations which are relatively unique for each object. Since most of the confusion cases are due to similar tactile features between objects, further experiments are performed by eliminating similar imprints based on crossed correlation measurement, boosting up the performance up to 7.87%.

Despite the success of the object recognition framework using the enhanced model of visual attention over randomly touching the object (blind touch), a more reliable way to

Table 1 – Classification accuracies for experiments conducted over six objects (adapted from [13]).

	Single tactile imprint					Single tactile imprints and probing locations		Multiple touches (dimensionality reduction and feature concatenation)	
	Enhanced visual attention				Blind touch, simulated data	Enhanced visual attention, simulated data	blind touch, simulated data	Enhanced visual attention, simulated data	
	All imprints		Similar imprints Eliminated						
	Simulated data	Real data	Simulated data	Real data				3 touches	4 touches
kNN	73.33%	51.60%	76.37%	58.97%	50.67%	84.86%	67.21%	99.36%	100%
SVM	67.78%	43.00%	75.65%	47.86%	48.89%	82.86%	56.12%	99.43%	100%

develop a highly reliable framework is to recognize objects from multiple touches. As such, an unsupervised network is first leveraged to reduce the dimensionality of features from each imprint and then multiple imprints are concatenated to be classified using kNN and SVM. Results confirm that, with at least four touches on the object at visually salient locations, the object can be perfectly recognized (Table 1).

Conclusion

Recent research makes use of biologically inspired computation and artificial intelligence as efficient means to solve real-world problems. Drawing inspiration from visuo-tactile collaboration in the human sensorial loop, this paper discusses possible approaches to exploit vision and touch sensing modalities to accelerate and optimize the process of 3D object representation and recognition for robotic manipulation.

References

[1] S. Kennett, M. Eimer, C. Spence, and J. Driver, "Tactile-visual links in exogenous spatial attention under different postures: convergent evidence from psychophysics and ERPs," *J. Cogn. Neurosci.*, vol. 13, no. 4, pp. 462-478, May 2001.

[2] S. Lacey and K. Sathian, "Visuo-haptic multisensory object recognition, categorization, and representation," *Front. Psychol.*, vol. 5, no. JUL, pp. 1-15, Jul. 2014.

[3] S. Foix, G. Alenya, and C. Torras, "Lock-in time-of-flight (ToF) cameras: a survey," *IEEE Sens. J.*, vol. 11, no. 9, pp. 1917-1926, Sep. 2011.

[4] E. Trucco and A. Verri, *Introductory Techniques for 3-D Computer Vision*. Upper Saddle River, NJ, USA: Prentice Hall, 1998.

[5] J. Geng, "Structured-light 3D surface imaging: a tutorial," *Adv. Opt. Photonics*, vol. 3, no. 2, p. 128, Jun. 2011.

[6] J. Davis, D. Neh, R. Ramamoorthi, and S. Rusinkiewicz, "Spacetime stereo: a unifying framework for depth from triangulation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 296-302, Feb. 2005.

[7] H. Yousef, M. Boukallel, and K. Althoefer, "Tactile sensing for dexterous in-hand manipulation in robotics—a review," *Sensors Actuators A Phys.*, vol. 167, no. 2, pp. 171-187, Jun. 2011.

[8] P. Regtien and E. Dertien, *Sensors for Mechatronics*. Cambridge, MA, USA: Elsevier, 2018.

[9] G. Rouhafzay and A.-M. Cretu, "A virtual tactile sensor with adjustable precision and size for object recognition," in *Proc. IEEE Int. Conf. Computational Intelligence and Virtual Environments for Measurement Syst. and Applications (CIVEMSA)*, 2018.

[10] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254-1259, 1998.

[11] G. Rouhafzay and A.-M. Cretu, "Perceptually improved 3D object representation based on guided adaptive weighting of feature channels of a visual-attention model," *3D Res.*, vol. 9, no. 3, p. 29, Sep. 2018.

[12] R. L. Klatzky and S. J. Lederman, "Identifying objects from a haptic glance," *Percept. Psychophys.*, vol. 57, no. 8, pp. 1111-1123, Nov. 1995.

[13] G. Rouhafzay and A.-M. Cretu, "Object recognition from haptic glance at visually salient locations," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 3, pp. 672-682, Mar. 2020.

Ghazal Rouhafzay (grouh050@uottawa.ca) is a Ph.D. degree candidate at the School of Electrical Engineering and Computer Science, University of Ottawa, Canada and a Research Assistant at the University of Québec in Outaouais, Canada. Her current research interest includes 3D sensing and modeling, biologically inspired visual and tactile data integration, deep learning techniques for visual and tactile object recognition.

Ana-Maria Cretu (M'10-SM'17) obtained her M.Sc. and Ph.D. degrees from the School of Electrical Engineering and Computer Science at University of Ottawa, Canada. She is currently Associate Professor with the Department of Computer Science and Engineering at the University of Québec in Outaouais, Canada. Her research interests include computational intelligence, soft computing, biologically-inspired computational models, tactile and vision sensing, and 3D object sensing, modeling and manipulation. She serves as Technical Committee Member for several international conferences and as a reviewer for journals and transactions.

Pierre Payeur is a Professor at the School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Canada. He is the Director of the Sensing and Machine Vision for Automation and Robotic Intelligence Research Laboratory and a Founding Member of the Vision, Imaging, Video and Autonomous Systems Research Laboratory. He holds a Ph.D. degree in electrical engineering from University Laval, Québec, Canada. His research interests include machine vision, tactile sensing, automation, robotics and computational intelligence.