SOM and LVQ classification of endovascular surgeons using motion-based metrics

Benjamin D. Kramer, Dylan P. Losey, and Marcia K. O'Malley

Rice University, Houston TX 77005, USA,
bkramer@rice.edu, dlosey@rice.edu, omalleym@rice.edu

Abstract. An increase in the prevalence of endovascular surgery requires a growing number of proficient surgeons. Current endovascular surgeon evaluation techniques are subjective and time-consuming; as a result, there is a demand for an objective and automated evaluation procedure. Leveraging reliable movement metrics and tool-tip data acquisition, we here use neural network techniques such as LVQs and SOMs to identify the mapping between surgeons' motion data and imposed rating scales. Using LVQs, only 50% testing accuracy was achieved. SOM visualization of this inadequate generalization, however, highlights limitations of the present rating scale and sheds light upon the differences between traditional skill groupings and neural network clusters. In particular, our SOM clustering both exhibits more truthful segmentation and demonstrates which metrics are most indicative of surgeon ability, providing an outline for more rigorous evaluation strategies.

Keywords: SOM, LVQ, skill assessment, surgical training

1 Introduction

Medical advancements in recent years have increased the popularity of endovascular surgery as an alternative to more traditional surgical methods [1]. In the most basic sense, endovascular surgery is a form of minimally invasive surgery (MIS) which allows access to various parts of the body through blood vessels and the endovascular system. The surgeon introduces a catheter into the vasculature of the patient, typically via the femoral artery, and from there navigates the catheter to the desired location so as to perform some type of procedure. During these procedures, surgeons must rely on fluoroscopy and other forms of medical imaging in order to determine tool position. This imaging is often limited, and complications may go unnoticed until they become too serious; therefore, it is imperative that surgeons be proficient at endovascular techniques. Aside from the risk of possible complications, surgeon skill level significantly affects clinical outcomes after successful surgeries [2].

1.1 Previous work

As a result, there is medical interest in understanding an effective means to determine a surgeon's skill [3]. There are presently two preeminent methods for

assessing a surgeon. The most common involves an expert observing task completion by a novice, which is entirely subjective and vulnerable to significant amounts of variability [4]. The second method is simply a measurement of the number of cases performed by the surgeon; although it stands to reason that an individual with more practice will likely be better, it is also likely that individual surgeons will improve at different rates. Either method is insufficient, and therefore a primary goal of the endovascular community is the development of an objective assessment technique [5], [6].

In an effort to more objectively study surgeons, sensors have been used to record the tool tip trajectory [7]. The results are then processed to calculate a variety of motion-based metrics; the most indicative of these metrics are correlated to user smoothness, such as minimum jerk [8] and spectral arc length [9]. An alternative, yet similarly-minded, method is the extraction of submovement number and duration from a larger task [10]. To date, researchers have attempted to show that there exist correlations between these movement metrics and the standard methods of skill evaluation. Surgeon force and motion signatures have been leveraged to objectively assess performance; hidden Markov models were then used to learn the nonlinear mapping between performance data and skill [11]. Lin et al. demonstrated the ability to decompose a surgical procedure into a series of sub-tasks by parsing raw motion data in order to provide on-line training feedback [12]. Estrada et al. specifically quantified the correlation between various metrics and the standard methods of surgeon evaluation on both manual and robotic platforms [13].

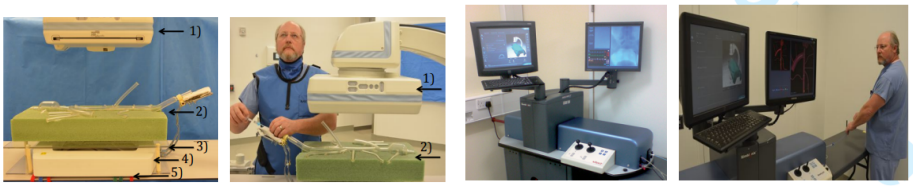
1.2 Motivation/Objective

Successfully mapping metrics to skill may improve training procedures, reduce the amount of oversight required, and ultimately automate this task. While the statistically significant correlation between various objective metrics and current subjective assessments is an important initial finding, it fails to provide a holistic approach to skill classification. Hence, the motivation for our work is to understand the mapping between movement metrics and surgeon proficiency, which we will reveal through neural networks. We will first train an LVQ to classify surgeons using standardized novice, intermediate, or expert labels, and then study the LVQ's accuracy using testing data. Next we will utilize SOMs to examine the underlying clusters; by comparing these SOM clusters with pre-labeled classes, we can evaluate the veracity of the medically imposed class labels. We hypothesize that the traditional "novice, intermediate, and expert" labeling—while commonly assumed to be correct—does not actually reflect the motion data, and, as such, more sophisticated classification is recommended. Our secondary goal is to identify which motion patterns contribute most to the surgeon's classification; this knowledge may improve the feedback which can be provided during and after the surgeon's training.

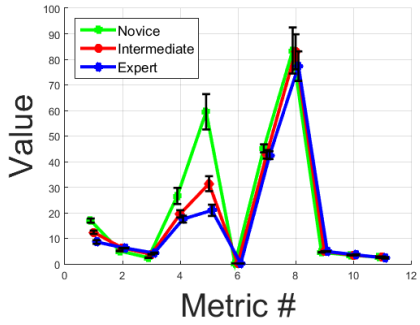
2 Methods

2.1 Input Data and Class Labels

The data used in this paper was collected during a previous study [13]. Actual and virtual tool-tip trajectories were recorded for fifteen surgeons over three sessions while completing four separate tasks. Five of the subjects (i.e., surgeons) were deemed "novices," six were labeled "intermediates," and the remaining four were regarded as "experts." The two platforms used during experimentation can be seen in Fig. 1, along with sample input vectors. For the purposes of our research, we did not differentiate between the platforms, sessions, or tasks, yielding a total of 120 separate trials. The motion metrics associated with each trial—described in more detail below—were then utilized as a unique input vector; hence, our results were obtained using 120 input vectors.



(a) Manual Simulator [13] (b) Virtual Simulator [13]



(c) Example Input Vector

Fig. 1: A comparison of the manual and virtual simulators which were navigated during the various tasks. In both platforms the surgeon is operating a catheter—in (a) the tip position is tracked using a magnet at the tool-tip, while (b) offers a platform leveraging teleoperation. In (c) each of the plotted points corresponds to one of the eleven motion metrics derived from the surgeon's trajectory. Average input vectors associated with a novice, intermediate, and expert surgeon are shown (standard error bars included).

The input vectors for our LVQ and SOM neural networks were constructed from previously calculated motion metrics. These metrics were all computed from the three-dimensional catheter position data, which was collected at 30 Hz frequency. Based upon the findings of Estrada et al. [13], we selected motion metrics which were shown to individually correlate with traditional skill labels. Eleven metrics (listed below) were chosen, and each comprised an element of the eleven-dimensional input vectors. Although the units for the various metrics are

not detailed here, it should be noted that they were kept consistent throughout our work. We found that our best results occurred with the inclusion of (1) Spectral Arc Length, (2, 3) Average Submovement Duration (LGNB and MinJerk Profiles), (4, 5) Number of Submovements (LGNB and MinJerk Profiles), (6) Normalized Velocity, (7) Mean Arrest Period Ratio (10% was used for this study), (8) Completion Time, (9, 10) Submovement Overlap (LGNB and MinJerk Profiles), and (11) Average Frequency. Example input vectors can be seen in Fig. 1. Note that these values are all well defined over a continuous range, and that the chosen metrics mitigated statistical outliers which may skew results of our input-space neural networks.

Each of these 120 input vectors was associated with a class label corresponding to the surgeon's proficiency; the three classes consisted of either "novice," "intermediate," or "expert." Forty input vectors were labeled novice, forty-eight input vectors were denoted intermediate, and thirty-two input vectors were termed expert. When performing supervised learning, stratified four-fold cross-validation was leveraged to select exclusive sets of ninety input vectors for training and thirty input vectors for testing.

2.2 Classification with LVQs

In order to determine the mapping from input data to desired classification, we used an LVQ with supervised learning [14]. More specifically, we used an LVQ2 with stratified four-fold cross-validation; the LVQ was initialized with 120 prototypes, as this was found to provide the best classification accuracy, where forty prototypes were allocated to novices, forty-eight were allocated to intermediates, and the remaining thirty-two were allocated for experts. Thus, this prototype allocation was done in proportion to class size. LVQ neurons were randomly initialized and scaled to the range of the input data. Ideally, the trained LVQ would adapt to the externally imposed classification structure, and, as such, would serve as an autonomous means towards identifying the class of the surgeon's skill—novice, intermediate, or expert—based solely on motion metrics. On the other hand, the reliability of the traditionally imposed class labels may be questionable [15]. These labels are based on the number of cases performed; however, it is conceivable that a surgeon could perform a large number of cases with improper technique, and therefore be labeled an "expert" by this traditional evaluation while actually maintaining a "novice" level of ability. In order to examine the performance of our classification with LVQs, we will show confusion matrix data and statistics across all four folds, as well as a visualization of our best results.

2.3 Clustering with SOMs

As we will demonstrate, the best classification accuracies obtained with LVQs were unsatisfactory, suggesting that more analysis into the label veracity is needed for effective machine learning. Further analysis of the input data—and,

in particular, clusters present in the input data—was performed and visualized through the use of SOMs [16]. We leveraged forty-nine prototypes for the SOM, which were arranged into a seven-by-seven rectangular grid in the lattice space. A Gaussian neighborhood function was used while updating the prototypes, and mU-matrix visualization was employed to visualize clusters. Our rationale for using an SOM was to capitalize upon the strengths of unsupervised learning; we sought to obtain an objective view of the data structure without needing potentially erroneous labels. Therefore, we had two primary goals behind this SOM application. First, we wanted to validate or disprove the classification labels (novice, intermediate, and expert) previously used for our LVQ training. By superimposing these labels over the SOM lattice while visualizing SOM clusters, we could test label veracity and hopefully understand why the LVQ machine learning underperformed. Second, we wanted to identify clusters within the data in order to determine the relative importance of surgeon attributes and motion methods when distinguishing between skilled and unskilled surgeons. By comparing the input vectors associated with different clusters, we can better understand which motion metrics were consistent and which varied amongst clusters. These insights may enable more efficient evaluation of surgeons and more directed training strategies. SOM clustering will be revealed through plots of the lattice space.

3 Results and Discussion

3.1 LVQ Classification Results

The results obtained by implementing an LVQ were reasonable, but did not provide sufficiently accurate classification for the purposes of automated evaluation. Our best results were obtained with an LVQ2 using a learning rate of 0.001 and 10,000 on-line learning steps, although other learning rates and learning step counts were tested. Both the training and testing accuracy were plotted as a function of learning steps to ensure that overtraining did not occur. To summarize, we consistently found that we were able to differentiate the skill groups and correctly classify surgeons within the novice, intermediate, and expert labels 80% of the time for training data and 50% of the time for testing data. In particular, the LVQ struggled to distinguish "intermediate" from "expert" surgeons, logically suggesting a larger skill gap from novice to intermediate than from intermediate to expert. This disparity is depicted in Fig. 2. We also note that, while LVQ1, LVQ2, and LVQ3 were tested, there was not significant variation among the performance of these algorithms.

By inspecting the confusion matrices, summarized in Fig. 3, we can further verify that novices were reasonably distinguished from intermediates and experts, but intermediates and experts were largely lumped together. We hypothesize that this stems from at least partially inaccurate training labels; the imposed classifications may not truly identify the skill level of each surgeon, since intermediate surgeons, despite having performed fewer cases than experts, may be

more proficient than their caseload suggests. Moreover, the use of only three classes is likely insufficient to accurately capture the gradient in surgeon skill, and perhaps more nuanced labels would better reflect our motion data. The overall statistics show that the LVQ procedure netted consistent and accurate training classification, but the testing accuracy and hence machine learning was unacceptable. We conclude that the LVQ was unable to generalize for the given data, and suggest that this inability stems from the lack of labeling precision and correctness for intermediate and expert surgeons. To verify this claim, we will subsequently explore SOM clusters in the data space.

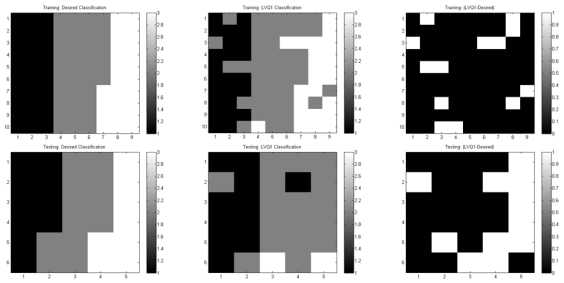


Fig. 2: Sample LVQ results. These plots are from one fold of the four-fold stratified cross-validation procedure: training classification top; testing classification bottom. The black pixels represent novice surgeons, the grey pixels represent intermediate surgeons, and the white pixels represent expert surgeons.

		Training Results			Testing Results		
		Novice	*Intermediate*	*Expert*	*Novice*	*Intermediate*	*Expert*
Labels	*Novice*	87.5% (26.25)	10% (3)	2.5% (0.75)	70% (7)	15% (1.5)	15% (1.5)
	Intermediate	6.25% (2.25)	81.9% (29.5)	11.8% (4.25)	20.8% (2.5)	52.1% (6.25)	27.1% (3.25)
	Expert	2.1% (0.5)	27.1% (6.5)	70.8% (17)	9.4% (0.75)	65.6% (5.25)	25% (2)

Classification Summary

	Training	Testing		Training	Testing
Mean	81% (291)	51% (61)	**Std.**	4.2% (15)	9.5% (11)

Fig. 3: Average confusion matrix over the four folds. Data is given in the form % of hits (number of hits). Diagonal elements represent correctly classified data, while off-diagonal elements show incorrect classifications. The mean and standard deviation for training and testing accuracy are also shown. Poor results likely stem from incorrect class labels, particularly between intermediates and experts.

3.2 SOM Clustering Results

Following the failure of LVQs to successfully identify this mapping, SOMs were applied to both test our concerns with the imposed classification labels and help us further explore nuances within the data. The best results presented in this paper were obtained using a seven-by-seven rectangular SOM grid in lattice space, where the forty-nine prototypes were initialized randomly over the input space. The learning rate α started at 0.005 and reached 0.001 following a linear

decrease across 100,000 learning steps; similarly, the Gaussian neighborhood width σ started at 4 and linearly decreased to 2 over the same number of learning steps. We experimentally observed the SOM training to converge after around 80,000 to 90,000 on-line learning steps, at which point no changes occurred in the mapping. The results shown below were found to be repeatable and superior to those identified using different parameters, which gives us confidence in the subsequent conclusions.

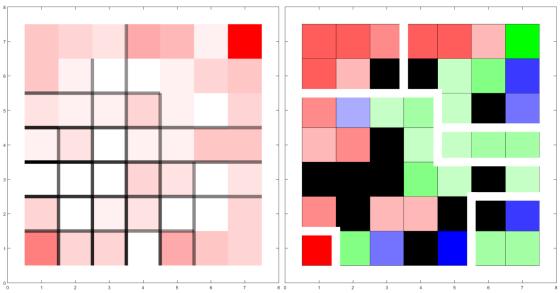


Fig. 4: SOM final results. The left visualization is a modified U-Matrix [17]; red-scale represents the number of mappings (i.e., relative density), while the gray-scale bars signify the distance between prototypes in the data space. The redder the neuron, the more input vectors are contained within its Voronoi cell; likewise, the darker the bar separating neurons, the greater the difference between their weight vectors. The right visualization shows the known surgeon classifications projected onto the SOM lattice—here red signifies novice, green represents intermediate, and blue indicates expert surgeons, with color intensity representing the number of mappings (more intensity again means increased density). Black neurons indicate that no input vectors are mapped to a particular node. The clusters found in the mU-matrix are identified using white lines in the right visualization. We can quickly observe that while novices (red) are primarily separated, clustering in the upper and lower left, intermediates (green) and experts (blue) are largely intermingled, clustering along the right side, a result which supports our LVQ findings.

Selecting the learning parameters as described above while observing the system visualizations depicted in Fig. 4, we repeatedly converged to a similar, if not the same, solution each time we trained the SOM. Instances in which we did not converge to the results outlined in Fig 4 involved some type of rotation of the lattice—however, this did not alter the SOM clustering. Using U-Matrix techniques, we readily discerned some distinct clusters which were identified by the SOM; we then checked these locations with superimposed novice, intermediate, and expert labels in the lattice space, and determined whether there existed agreement between medically defined clusters and clusters identified by the SOM.

From the modified U-Matrix density map and the projection of classifications into lattice space, we can deduce (a) that there exist some SOM clusters which roughly correspond with traditional groups, but (b) other SOM clusters disagree with the medical consensus. We have marked these SOM identified clusters in Fig. 5. For instance, the bottom left section of the SOM lattice clearly clusters several surgeons who performed poorly, and are correctly labeled as novices. Likewise, the top left SOM cluster corresponds to another group of novice surgeons, which again matches the medical labeling. Moving to the right side of the SOM lattice, however, we can see two regions: in the upper right, there exists

a mixed cluster—some experts, intermediates, and novices are included here, suggesting labeling inaccuracy. Finally, in the bottom right of the SOM lattice we find a cluster of increasing ability, with intermediates and experts grouped together; perhaps these surgeons are closer in ability than their classification would suggest. By applying SOMs to the input space of motion metrics, we were therefore able to demonstrate that a surgeon's experience is not sufficient when attempting to classify that surgeon's skill. Although there are some similarities between the medical labels and SOM clusters, there is also sufficient disparity to suggest that perhaps more precise skill assessment is required. These findings also explain the inability of our LVQs to distinguish "intermediate" and "expert" surgeons, as SOM clusters revealed overlaps between these classifications.

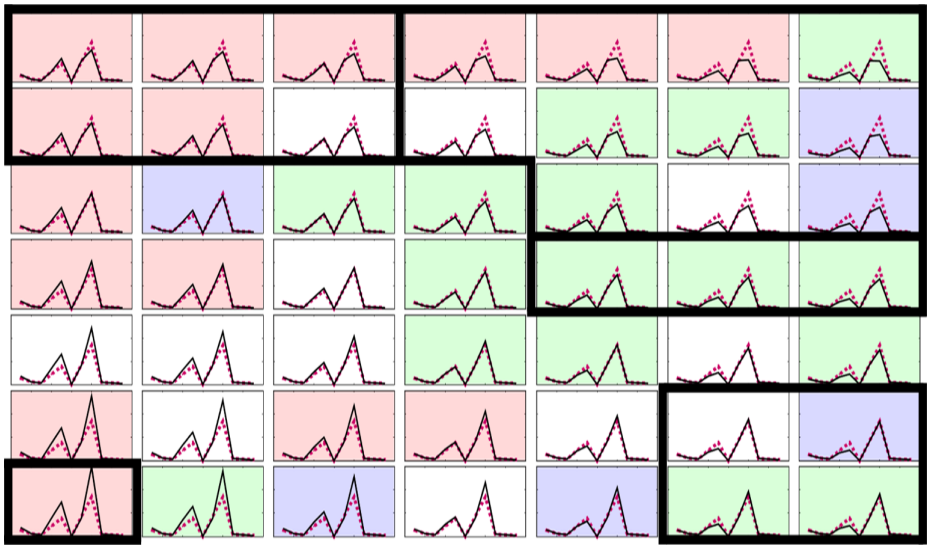


Fig. 5: SOM weight vector plotted in the grid cells. This figure shows both the final results of our SOM grid with the known classifications projected onto the lattice, as well as the weight vector of each PE with respect to the average weight vector across all nodes. The weight vector of a given PE is shown in black, while the average weight vector across all nodes is plotted in a dotted magenta line. The color coding of the prototypes is the same as before, with a slight fading of the colors in order to better visualize the weight vectors. Black boxes were used to mark the SOM cluster boundaries identified in 4. By comparing the differences in weight vectors between members of different clusters, we can visualize which metrics most impact distinctions in surgeon skill. With respect to the average weight vector, novices appear to complete the task in less time but require an increased number of motions; on the other hand, proficient surgeons move slowly but smoothly, reducing submovement duration and number. The combination of SOM clustering and neuron weight vectors reveals errors within traditional labeling and provides insight into important motion attributes. The existing labeling of novice, intermediate, and expert does not agree with knowledge gained through motion metrics (as shown by differences in clustering), and the contribution of various metrics can be analyzed to yield better categorization (as shown by comparing weight vectors).

In order to further investigate clustering and the distinctions between various groups, it was instructive to look at the weight vector within these individual clusters, as illustrated in Fig. 5. There are a few hypotheses which can be formed from visualizing these prototypes and clusters. First, completion time is not necessarily an accurate measure of skill. In fact, completion time appears

to be somewhat counter-intuitive; experts often take longer than less successful intermediates and novices, perhaps because they are utilizing slower and more deliberate movements. A quick procedure is ideal, but not if it comes at the cost of deliberate, precise movements. Second, some metrics may provide redundant differentiation, therefore requiring the use of fewer metrics—and other metrics may be entirely irrelevant for classification purposes. Finally, the number of submovements appears to be particularly useful when distinguishing surgeons; we observed that the most proficient cluster employed substantially smoother motions than did novice or mixed clusters.

With these ideas in mind, we can describe the five classes of surgeons from SOM clustering (Fig.5). Class one (lower left) will perform the task slowly and with very little smoothness; likely true beginners. Class two (upper left) will perform the surgery quickly with little smoothness; likely novice surgeons. Class three (upper right) will perform the surgery quickly at the expense of some smoothness metrics; likely competent surgeons primarily concerned with completion time. Class four (middle right) will perform the surgery above average in terms of time and smoothness; likely experienced surgeons. Class five (lower right) will perform the surgery at an average pace with exceptional dexterity; likely skilled, precise surgeons.

4 Conclusions and Future Work

Based on the results of our LVQ and SOM, there does appear to be some consistent mapping between motion metrics and desired classification; using the LVQ we achieved around 50% testing accuracy. We hypothesized that this poor LVQ machine learning, particularly when discerning between intermediate and expert surgeons, stemmed from inaccurate class labeling. Using the SOM approach, we were able to identify some clusters which roughly corresponded to the known classification groups; however, we also discovered that several clusters disagreed with the given labels. Indeed, from Fig. 4 we were able to conclude that the traditional labeling based on surgeon experience disagreed with SOM clustering in the motion metrics. We were further able to suggest which metrics may best be able to indicate ability, as can be seen in Fig. 5. By replacing the subjective medical grouping with the actual measured features, we may be able to improve on skill assessment for endovascular surgeons. Similar to the work by Cotin et al. [5], we suggest that it may be better to first identify statistics which are significant to expert clusters, and then create a scoring system which classifies users based on their accordance with those statistics. Summarily, SOM clustering, as seen in Figure 5, helps accomplish our goals of both disproving classical labels and suggesting improved alternatives.

References

1. A. Schanzer, R. Steppacher, M. Eslami, E. Arous, L. Messina, and M. Belkin, "Vascular surgery training trends from 2001-2007: A substantial increase in total

procedure volume is driven by escalating endovascular procedure volume and stable open procedure," *Journal of Vascular Surgery*, vol. 49, no. 5, pp. 1399–1344, 2009.
2. M. Cox, D. M. Irby, R. K. Reznick, and H. MacRae, "Teaching surgical skills—changes in the wind," *New England Journal of Medicine*, vol. 355, no. 25, pp. 2664–2669, 2006.
3. C. E. Reiley, H. C. Lin, D. D. Yuh, and G. D. Hager, "Review of methods for objective surgical skill evaluation," *Surgical Endoscopy*, vol. 25, no. 2, pp. 356–366, 2011.
4. A. Darzi and S. Mackay, "Assessment of surgical competence," *Quality in Health Care*, vol. 10, no. suppl 2, pp. ii64–ii69, 2001.
5. S. Cotin, N. Stylopoulos, M. Ottensmeyer, P. Neumann, D. Rattner, and S. Dawson, "Metrics for laparoscopic skills trainers: the weakest link!" in *Medical Image Computing and Computer-Assisted Intervention*. Springer, 2002, pp. 35–43.
6. G. J. M. Van Hove, P. D. andTuijthof, E. G. G. Verdaasdonk, L. P. S. Stassen, and J. Dankelman, "Objective assessment of technical surgical skills," *British Journal of Surgery*, vol. 97, no. 7, pp. 972–987, 2010.
7. J. Kuipers, "Object tracking and determining orientation of object using coordinate transformation means, system and process," Patent, 1975.
8. N. Hogan and D. Sternad, "Sensitivity of smoothness measures to movement duration, amplitude, and arrests," *Journal of Motor Behavior*, vol. 41, no. 6, pp. 529–534, 2009.
9. S. Balasubramanian, A. Melendez-Calderon, and E. Burdet, "A robust and sensitive metric for quantifying movement smoothness," *IEEE Trans. on Biomedical Engineering*, vol. 59, no. 8, pp. 2126–2136, 2012.
10. B. Rohrer and N. Hogan, "Avoiding spurious submovement decompositions: A scattershot algorithm," *Biological Cybernetics*, vol. 94, no. 5, pp. 409–414, 2006.
11. H. Rafii-Tari, C. J. Payne, J. Liu, C. Riga, C. Bicknell, and G.-Z. Yang, "Towards automated surgical skill evaluation of endovascular catheterization tasks based on force and motion signatures," in *Proc. IEEE Int. Conf. on Robotics and Automation*, 2015, pp. 1789–1794.
12. H. C. Lin, I. Shafran, D. Yuh, and G. D. Hager, "Towards automatic skill evaluation: detection and segmentation of robot-assisted surgical motions," *Computer Aided Surgery*, vol. 11, no. 5, pp. 220–230, 2006.
13. S. Estrada, M. K. O'Malley, C. Duran, D. Schulz, and J. Bismuth, "On the development of objective metrics for surgical skills evaluation based on tool motion," in *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*. IEEE, 2014, pp. 3144–3149.
14. T. Kohonen, "Learning vector quantization," in *Self-Organizing Maps*, ser. Springer Series in Information Sciences. Springer, 1995, vol. 30, pp. 175–189.
15. J. Bismuth, M. A. Donovan, M. K. O'Malley, H. F. El Sayed, J. J. Naoum, E. K. Peden, M. G. Davies, and A. B. Lumsden, "Incorporating simulation in vascular surgery education," *Journal of Vascular Surgery*, vol. 52, no. 4, pp. 1072–1080, 2010.
16. T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, no. 1, pp. 1–6, 1998.
17. E. Merényi, K. Tasdemir, and L. Zhang, "Learning highly structured manifolds: harnessing the power of SOMs," in *"Similarity-based clustering," Lecture Notes in Computer Science*. Springer, 2009, pp. 138–168.