Transferring Dexterous Surgical Skill Knowledge between Robots for Semi-autonomous Teleoperation

Md Masudur Rahman[1*], Natalia Sanchez-Tamayo[2*], Glebys Gonzalez[2], Mridul Agarwal[3], Vaneet Aggarwal[2,3], Richard M. Voyles[4], Yexiang Xue[1], and Juan Wachs[2]

Abstract— **In the future, deployable, teleoperated surgical robots can save the lives of critically injured patients in battlefield environments. These robotic systems will need to have autonomous capabilities to take over during communication delays and unexpected environmental conditions during critical phases of the procedure. Understanding and predicting the next surgical actions (referred as "surgemes") is essential for autonomous surgery. Most approaches for surgeme recognition cannot cope with the high variability associated with austere environments and thereby cannot "transfer" well to field robotics. We propose a methodology that uses compact image representations with kinematic features for surgeme recognition in the DESK dataset. This dataset offers samples for surgical procedures over different robotic platforms with a high variability in the setup. We performed surgeme classification in two setups: 1) No transfer, 2) Transfer from a simulated scenario to two real deployable robots. Then, the results were compared with recognition accuracies using only kinematic data with the same experimental setup. The results show that our approach improves the recognition performance over kinematic data across different domains. The proposed approach produced a transfer accuracy gain up to 20% between the simulated and the real robot, and up to 31% between the simulated robot and a different robot. A transfer accuracy gain was observed for all cases, even those already above 90%.**

I. Introduction

There is an increasing interest in using teleoperated surgical robots for austere environments (such as directly on the battlefield) since they can provide timely interventions to patients with life-threatening injuries [1]. These systems are sensitive to delays intrinsic to the limited bandwidth of many austere environments [2], so there is a need for platforms with semi-autonomous capabilities that can assist the surgeon (or medic) when communication is hindered. These systems require a high level understanding of their state and environment to take over when required. In order to effectively interpret the environment, it is helpful to recognize the current and previous surgical actions, referred to as 'surgemes' [3], that are being preformed by the surgeon. In fact, the accurate recognition of surgeme-like primitives is valuable for many robot application domains beyond surgery, such as manufacturing [4] or waste handling [5].

The use of machine learning approaches for surgeme recognition requires collecting a substantial amount of data, which is challenging to obtain in austere settings [6]. Alternatively, it would be desirable to leverage the abundance of data available from more accessible environments to find recurrent patterns and apply such insights to new scenarios [7]–[9]. Hence, the knowledge learned from numerous accessible platforms could be transferred to deployable robotic platforms in less hospitable environments. However, field medical robots are diverse, holding different kinematic configurations, workspaces, and operate under partially unknown constraints. Such domain differences could hamper state-of-the-art approaches and prevent models learned on one platform to generalize across other platforms of disparate morphologies [10]. Addressing such challenges involves coming up with an effective transfer learning architecture than can generalize over different robots and surgical settings.

This paper presents an approach for identifying surgemes performed by a surgeon through teleoperation that can deal with variable environments and minor shifts in robot hardware. The proposed machine learning architecture leverages both images and kinematic features to transfer knowledge obtained from a surgical simulator to real deployable robots. The features were designed to be applicable to any robotic platform and surgical setup without the need of environment-specific modeling. The architecture has been tested on a surgeme classification task under two scenarios: (1) the source and target data are from the same domain (e.g. same robot) and (2) the source and target data come from different domains (e.g. different robots). The use of robot kinematic and visual data to improve knowledge transfer between different domains is a major enhancement over the current state-of-the-art. The contributions of this paper can be summarized as follows: i) a surgeme classification architecture has been developed for settings with high variability; ii) the impact of adding visual information to kinematic data for surgeme recognition during a transfer scenario is demonstrated.

The rest of this work is presented as follows: Section II discusses the prior work. Section III gives an overview of the robotic dataset used. Section IV describes the methods used for surgeme classification. Section V shows the experimental setup and results. Section VI presents discussion, and Section VII concludes the paper with a discussion on future work.

[*] These authors contributed equally to this work

[1] Department of Computer Science, Purdue University, West Lafayette, IN, 47907, USA rahman64, yexiang@purdue.edu

[2] School of Industrial Engineering, Purdue University, West Lafayette, IN, 47907, USA sanch174, gonza337, vaneet, jpwachs@purdue.edu

[3] School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, 47907, USA agarw180@purdue.edu

[4] School of Engineering Technology, Purdue University, West Lafayette, IN, 47907, USA rvoyles@purdue.edu

II. Background and related work

Surgical skill modeling [11] consists of decomposing surgical tasks into sets of finite, well defined and quantifiable maneuvers. Such maneuvers can be used to create datasets of surgical skills, and are referred to as surgemes [12]. Surgeme analysis has been used to model skill, to classify tasks, and to assess proficiency [13]. Furthermore, surgeme recognition provides the opportunity for feedback during surgical training [14] [15]. Finally, surgeme classification has been identified as a key step in robot automation [13] and is transferable to other domains of robot autonomy.

Commonly adopted approaches for surgeme recognition operate either on the user space or task space. In the task space, robotic kinematic data have been matched to templates using Hidden Markov Models (HMM), where the surgemes correspond to one or more states of the HMM [16], [17], [18]. Ahmidi *et al.* also used kinematics to compare benchmarks on approaches for joint segmentation and classification (sparse HMM, Markov semi-Markov conditional random field, and skip-chain conditional random field) with feature-based approaches such as bag of spatio-temporal features and linear dynamical systems [19]. Other approaches that use structured prediction include the use of a Skip-Chain Conditional Random Field (SC-CRF) [20], and a Latent Convolutional Skip-Chain Conditional Random Field (LC-SC-CRF) [21].

The previously mentioned works were developed using the JIGSAW dataset (JHU-ISI Gesture and Skill Assessment Working Set) [7]. The JIGSAW dataset comprises synchronized video and kinematic information for three procedures preformed with the da Vinci Surgical System on a benchtop model: suturing, knot tying, and needle passing. DiPietro et al. used a recurrent neural network to classify kinematic data for activity recognition based on the JIGSAW dataset for surgeme classification and the MISTIC-SL dataset for longer segments called maneuvers [9].

A limiting factor of the JIGSAW dataset is that it does not contemplate variability in the environment or domain, keeping a constant configuration of objects of interest, the initial conditions of the task, and a single robot (the 'da-Vinci'). The lack of variability in task conditions hinders generalization and the capacity of transfer learning. In contrast, Madapana *et al.* [8] used a transfer learning approach based on models trained with kinematic data from a new dataset called the DExterous Surgical SKill (DESK) dataset. Using this dataset, the authors trained a model on a simulated robot and applied it to different real robots. The approach obtained 55% accuracy in surgeme recognition over a teleoperated surgical robot when trained only over simulation data and achieved an improvement of 34% when a small percentage of real data was added to the training set. The approach demonstrated the power of transfer learning to augment training in austere environments. Nonetheless, this approach was based solely on kinematic information from the robot and did not consider information related to the state of the environment.

The DESK dataset is challenging because it offers high variability by randomizing the object locations and initial conditions of the tasks over three different robotic domains (see Fig. 1). For applications with high variation in object placement and appearance, image information is of paramount importance, as it facilitates obtaining features related to changes in the scene that the kinematic information of the robot does not account for. Previous work on surgeme segmentation and classification has included image information, yielding equally good results as the kinematic approaches. These approaches demonstrate that both kinematics and video data capture relevant information for activity classification. The work in [22] modelled each surgeme video clip as the output of a linear dynamical system, extracting spatio-temporal features from each video to learn a bag-of-features model. Other methods that employ both modalities (e.g. kinematic and visual) have increased performance of surgeme segmentation and recognition [18], achieving better results than the structured learning approaches that use kinematic data alone. However, these methods did not evaluate the combination of kinematic and visual data to improve transfer learning across domains. The addition of image information allows to address challenging and diverse settings where visual cues might be the unique information available about the state of the environment.

We propose an architecture for surgeme classification across different domains, that uses a compact image representation with kinematic data. Our system is tested in a transfer leaning scenario where models are trained using simulation data and tested with two real robots. Section IV describes the implementation details of the architecture.

III. Dataset description

The dataset used in this paper is the DESK dataset described in [8][1]. This dataset provides synchronized RGB images, depth and kinematic information for the peg transfer task from multiple domains including two real robots (Taurus II and YuMi) and a simulation environment (Taurus II), as shown in Figure 1. The DESK dataset attempts to account for the complexity of transfer learning between dissimilar robots by introducing intentional variance in peg board configuration and object size and appearance.

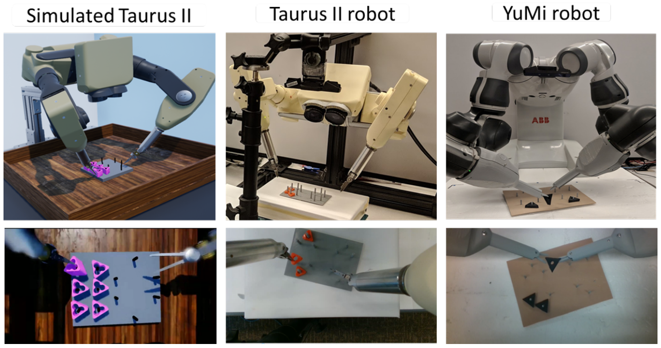

Fig. 1: Robotic system setup for peg transfer in the DESK dataset

[1] Available at https://github.com/nmadapan/Forward_Project

The RGB video, the depth video and the kinematic data are segmented according to surgemes observed in RGB video frames. Additional variability is added to the dataset by randomizing the pick and place locations for the pegs and the orientation of the board, while leaving the order of the pegs to be transferred unrestrained. In addition, successful and failed surgemes are included in the dataset along with the subsequent recovery maneuvers. From the DESK dataset, we used Taurus simulator S1-S5 [2] , Taurus S1-S8, and YuMi S1-S8. Details of the surgeme statistics can be found in the Table I. The dataset includes kinematics variables that represent robot position, orientation and gripper status as shown in the Table II.

TABLE I: Surgemes in the peg transfer task. The columns indicate surgeme ID, name of the surgeme, number of instances present for each surgeme for the simulator, real Taurus and the YuMi robot.

ID	Surgeme name	# Sim	# Taurus	# YuMi
S1	Approach peg	90	110	117
S2	Align & grasp	92	111	123
S3	Lift peg	91	111	123
S4	Transfer peg - Get together	84	111	117
S5	Transfer peg - Exchange	80	111	118
S6	Approach pole	76	109	117
S7	Align & place	75	107	116

TABLE II: Kinematic variables. Note that ts is the Unix timestamp, $\vec{J}$ is the vector of joint angles, $\vec{p}$ is the position vector (x, y and z), $\vec{\theta}$ be the Euler angles (yaw, pitch and roll), gs is the gripper state of the end-effector and R be the 3 x 3 rotation matrix. (adopted from DESK [8])

Taurus		**Taurus simulator**		**YuMi**	
ID	Variable	ID	Variable	ID	YuMi
1	ts	1	ts	1	ts
2-13	R and $\vec{p}$	2-4	$\vec{p}$	2-8	$\vec{J}$
	-	5-7	$\vec{\theta}$	9-11	$\vec{p}$
14-16	$\vec{p}$	8-14	$\vec{J}$	12-20	R
17	gs	15	gs	21	gs

IV. Methodology

Our approach combines features extracted from video image and kinematic data to perform transfer learning in surgeme recognition. The approach is based on robot-agnostic feature extraction that can be applied to different environments in a transfer learning scenario.

An overview of our system architecture is given in figure 2. The proposed approach uses a pretrained Convolutional Neural Network (CNN) to extract meaningful features from video images. The images extracted from the videos were resized from 1920×1080 to a 228×128. The experiments in this paper are preformed using the CNN model ResNet18 [23] pre-trained on ImageNet. The last two layers (i.e., fully connected and softmax) of ResNet18 were removed to create a high dimensional feature representation for each video frame. The extracted features have dimensions of $1 \times 512 \times 4 \times 8$, which were flattened to a single vector of size $N = 16384$. Due to the large dimension in the output of the ResNet18 model a *dimensional reduction module* was required to reduce each image/frame features to a lower dimensional vector ($M = 30$, where $M << N$). Principal Component Analysis (PCA) was used to reduce the dimension of each frame image to a lower 30-dimensional vector. PCA is a statistical method that applies an orthogonal transformation to convert a set of observations into a set of uncorrelated variables. Intuitively, this transformation extracts components (i.e., features) with the largest possible variance and thus helps classifiers to distinguish between different surgeme classes.

In parallel, we extracted features from kinematic data using a similar approach to the one proposed in the DESK [8], by reducing the kinematic features from multiple domains to the commonly shared features: position, orientation and gripper status of the end-effector (14 features: seven features in each arm). Since each surgeme instance consists of a variable number of frames, we re-sampled (via linear interpolation) the original instances to a fixed number of frames (40) to generate a sequential feature instance for each surgeme. In particular, for the kinematics, we concatenated the 14 features corresponding to each frame, to create a single 560 dimensional vector per surgeme ($40 \times 7 \times 2$). On the other hand, for the video data, we concatenated the 30 features obtained after the PCA reduction into a single feature vector of 1200 dimensions (40×30) that represents each surgeme. Another PCA reduction was applied on the 1200-dimensional feature vector to obtain a 100-dimensional feature vector for entire image sequence of 40 frames.

In our pipeline, visual features and kinematic features came from different distributions. Thus, to take advantage of both visual and kinematic modalities the supervised learning algorithms were trained separately and their output class probabilities were combined as shown in the following equation:

$$P(C) = \lambda P_{kin}(C) + (1 - \lambda) P_{video}(C), \qquad (1)$$

where C is the class and $\{P_{kin}(C),\ P_{video}(C)\}$ are the probabilities given by supervised learning models using kinematic and video features respectively. Here, λ is the hyper-parameter (from 0 to 1), which modifies the contribution of the models based on kinematics or video frames. When λ is set to 0, the classification is based on video data alone. If $\lambda =$ 1, the classification uses only kinematic data. Note that the architecture allows to plug in different supervised machine learning models for kinematics and video. Thus, the models for video and kinematics can be designed separately. For simplicity, in our experiments, the same supervised learning model for both video and kinematic features was used.

The final feature set consists of the gripper's position, orientation, status (open/closed) and image features extracted from the pre-trained ResNet18. These features are agnostic to the robot. Hence, the same set can be used even when the task or the robot varies. Having a common set of features facilitates surgeme classification across different domains, since it helps models to leverage on the information coming from another domain.

[2] a subset from 8 mobile pegboard subjects [8]

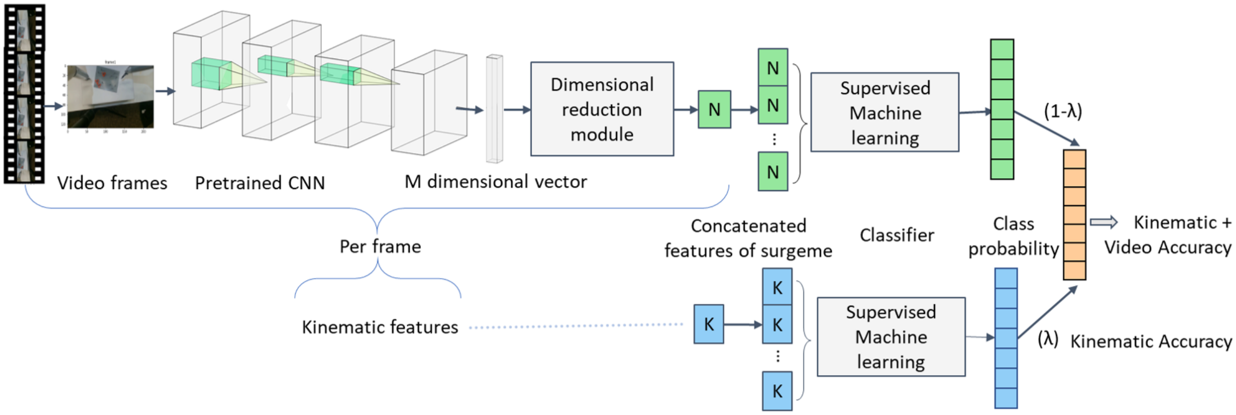
Video frames
Pretrained CNN
M dimensional vector
Dimensional reduction module
N
N
N
N
Supervised Machine learning
(1-λ)
Kinematic + Video Accuracy
Per frame
Concatenated features of surgeme
Classifier
Class probability
Kinematic features
K
K
K
K
Supervised Machine learning
(λ)
Kinematic Accuracy

Fig. 2: Architecture overview of the approach, M: Number of dimensional features extracted using pre-trained CNN, K: number of kinematic features, N: number of image features after Dimensional reduction module.

V. Experiments and results

Experimental Setup: The proposed architecture was tested over two scenarios: no-transfer scenario (train and test data are obtained from the same domain), and domain-transfer scenario (train on one domain and test on the other). The complete domain transfer scenario used the Taurus simulator (S) data or a combination of simulation and real robot data (S+R) during training, while the algorithm was tested entirely on the real robot data. A value α represents the percentage of real data (i.e., target domain) that was added to the simulator data (source domain) for the model training. The rest of the real data (in the target domain) was used as the testing set. Thus, $\alpha = 0$ indicates a complete transfer.

Two supervised learning methods were tested: (1) Random Forest (RF), and (2) Support Vector Machines (SVM), using the scikit-learn [24] implementation. A five-fold cross-validation approach was used with a data split of 80-20% for training and testing respectively. Furthermore, *hyperparameter setting*, we used a $kernel = poly$ for SVM classifier. For RF, we set $n\_estimators = 200$ (number of trees in the forest), and maximum depth = 10. We set the combined model's hyper parameter λ to 0.5 for RF and to 0.8 for SVM (empirical best).

Results: Table III shows the surgeme recognition results using our approach for kinematics, video information and a combination of both modalities for the non-transfer scenario. Kinematic (Kin) features alone perform substantially better than features based on video frames (Visual). RF performs better than SVM for visual data achieving up to 63% accuracy on the Taurus real robot. Moreover, in most of the cases, accuracies improved when both kinematic data and image features are used (Kin+Visual). In general, adding visual data improved the classification accuracy of surgemes compared to the use of kinematic or visual data independently.

Two transfer learning scenarios were evaluated: (i) Taurus simulator to Taurus real robot (Sim2Taurus) and (ii) Taurus simulator to YuMi (Sim2YuMi). The accuracy results for the domain-transfer scenario: simulator + real-robot (S+R) are presented on Figure 3. The results show that kinematic+visual features surpass the accuracy and consistently outperforms results that use only kinematic features.

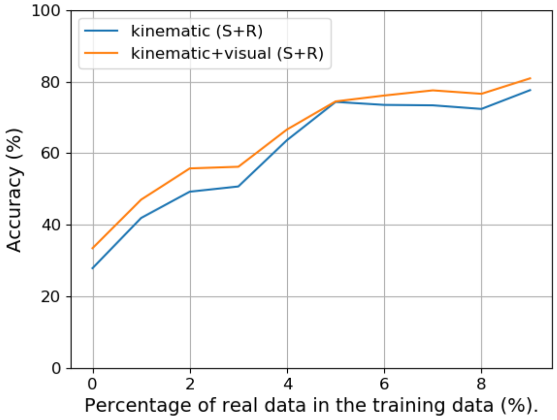

Fig. 3: Performance comparison for transfer learning Sim2Taurus (Random Forest). Overall transfer accuracy improved quickly and it reaches over 93% with only 50% real data in the training. Details in Table IV.

Figure 4 shows that the performance of both the (Visual) and (Kin+Visual) data increases as we add more real-robot data (YuMi) in the training (S+R). The improvement obtained for the transfer learning scenario using (Kin+Visual) data is higher than the knowledge transfer based on kinematic data alone. Hence, when training data is limited on the target domain, the addition of visual data from simulation could help boost initial classification accuracy.

Table IV shows transfer accuracy for kinematic only features (Kin)[3], and both kinematic and visual features (Kin+Visual). It also shows a percentage gain which is calculated as

$$Gain = \frac{A_2 - A_1}{A_1} \times 100, \tag{2}$$

where A_1 and A_2 are the accuracies using Kin features, and Kin+Visual features, respectively. It can be observed that adding visual features boosts the transfer accuracy in both Sim2Taurus and Sim2YuMi robot settings achieving up

[3] Note that this paper used a subset of simulated data, and thus the kinematic only transfer results are slightly different than that reported in [8]. For instance, RF accuracy from Sim2Taurus at $\alpha = 0$ is 28% as compared to previously reported 34%.

TABLE III: Classification accuracy on the no-transfer scenario using RF and SVM on kinematic only (Kin), visual only (Visual), and both (Kin+Visual) features. In most cases adding visual features with kinematic features improve the accuracy.

	RF			SVM		
Robot	Kin	Visual	Kin+Visual	Kin	Visual	Kin+Visual
Simulator (Taurus)	92.69	56.46	93.03	91.66	33.16	92.51
Taurus	94.68	63.77	96.49	91.43	35.19	92.73
YuMi	96.87	60.65	97.23	92.90	21.91	93.38

TABLE IV: Domain transfer accuracy (Random Forest) when the models are trained on the Taurus simulator robot and tested on real robots (Taurus (Sim2Taurus) and YuMi (Sim2Yumi)). α is the percentage of real-robot data used in the training. Features: kinematic only (Kin), visual only (Visual), and both (Kin+Visual).

	Sim2Taurus			Sim2Yumi		
α(%)	Kin	Kin+Visual	Gain(%)	Kin	Kin+Visual	Gain(%)
0	27.92	33.51	20.00	14.20	14.80	4.24
1	41.94	47.05	12.19	25.64	33.78	31.75
2	49.27	55.22	13.17	37.42	43.07	15.08
3	50.74	56.22	10.82	39.03	46.84	20.00
4	63.65	66.62	4.67	50.38	54.14	7.46
10	78.93	82.54	4.57	74.73	73.40	-1.78
50	93.77	94.03	0.28	88.70	90.14	1.62
90	93.51	96.10	2.77	95.24	96.43	1.25

to 20% and 31% gain respectively. Another observation is that the performance improvement is generally higher for smaller α values. Intuitively, when the target domain data is limited during training, visual information can play an important role in increasing initial transfer accuracy. As more target domain data is added to the training, both Kin and Kin+Visual accuracy increase quickly. In particular, when only 50% data is added to the simulator training, the accuracy reaches over 93% and 88% for Sim2Taurus and Sim2YuMi respectively.

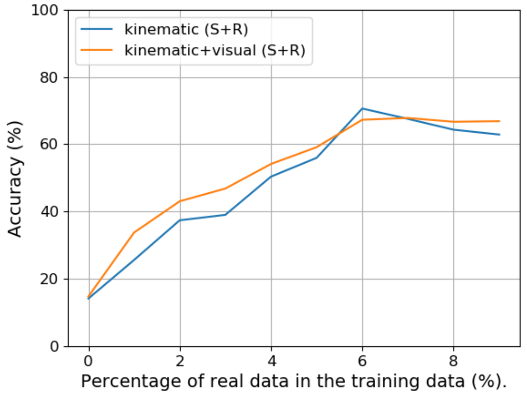

Fig. 4: Performance comparison for transfer learning from Sim2YuMi (Random Forest). Overall transfer accuracy improved quickly and it reaches over 88% with only 50% real data in the training. Details in Table IV.

VI. Discussion

As shown in Section V, higher recognition was generally achieved when the predicted label is obtained from both kinematics and visual information. Even though recognition accuracy was poorer when only visual features were used for the classification model, adding visual to kinematic data incremented recognition performance. Particularly, in the experiment involving transfer learning across domains, visual information complemented and improved recognition in transfer scenarios, especially in cases where there was little to no available training information on the target domain. In the no-transfer scenario, the recognition accuracy obtained using only kinematic features was already above 90%, so when visual features scores were included in the prediction, results were only marginally higher.

Image features encode information about the motion of the robot along with information about the scene, environment and relevant objects. Thus, the image features can provide knowledge about the aspects of the task that the kinematics cannot account for. However, the encoded features of the image provide a noisier representation for robot motion. Since these features are obtained from a 2D image, which is a projection of the 3D world, the input is already reducing the dimensionality of the spatial data. In contrast, the kinematic data is working at millimeter precision while the image that was fed to the network represents the robot's space at the pixel level.

The extracted kinematic features make no assumptions on the arm's morphology. Thus, the approach can be applied to different robotic systems facilitating a transfer learning scenario. Additionally, the image features extraction module of our architecture is not constrained to a particular CNN model. Thus the ResNet18 can be replaced by other preferred CNN based models.

Although image features extracted from a pre-trained CNN network can be useful to extract meaningful features from domains with limited training data, they fall short when representing the common traits of objects of interest across domains. This is currently a limitation of this work, since the features that are obtained from the ResNet18 are not particularly designed for surgical tasks. Leveraging these commonalities across domains, beyond the differences in environment appearance, might provide information particularly helpful for the transfer learning scenario. For example, between the simulated and the real robot environment, the appearance of the peg is different although they share the same role in the task.

VII. Conclusions and future work

This paper proposed a systematic way to incorporate visual information to improve surgeme classification in surgical robotic tasks. Several experiments were conducted on datasets from three robotic domains. Results show that the visual features improve the performance of surgeme classification when used along with kinematic data. Currently the features are combined using late fusion. Future work can explore early fusion techniques to improve the classification accuracy.

In the transfer scenario, the proposed approach improved the knowledge transfer accuracy between domains by leveraging visual features effectively with the kinematic features while requiring little data from the target domain. In particular, the visual data boosted the transfer accuracy between simulator robot (Taurus) and a different robot (YuMi) by up to 31% when little real data from the target was added with the simulator data for training. For the case of transfer between Taurus simulator to Taurus real robot, an increase of up to 20% was achieved when incorporating visual data to the model.

Acknowledgments

This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs under Award No. W81XWH-18-1-0769 and by the NSF RoSe-HUB Center under CNS-1439717. The authors would like to acknowledge Jyothsna Padmakumar Bindu for her feedback during the preparation of this manuscript. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the funders.

References

[1] P. Garcia, J. Rosen, C. Kapoor, M. Noakes, G. Elbert, M. Treat, T. Ganous, M. Hanson, J. Manak, C. Hasser, *et al.*, "Trauma pod: a semi-automated telerobotic surgical system," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 5, no. 2, pp. 136–146, 2009.

[2] J. Loyall, M. Gillen, J. Cleveland, K. Usbeck, J. Sterling, R. Newkirk, and R. Kohler, "Information ubiquity in austere locations," *Procedia Computer Science*, vol. 10, pp. 170–178, 2012.

[3] 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, pp. 220–230, Jan. 2006.

[4] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and Autonomous Systems*, vol. 57, no. 5, pp. 469 – 483, 2009.

[5] T. Soratana, M. V. S. M. Balakuntala, P. Abbaraju, R. Voyles, J. Wachs, and M. Mahoor, "Glovebox handling of high-consequence materials with super baxter and gesture-based programming - 18598," in *Waste Management (WM 2018), 44th International Symposium on*, Waste Management, March 2018.

[6] S. H. Y. John M. Fossaceca, "Artificial intelligence and machine learning for future army applications," 2018.

[7] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. Béjar, and D. D. Yuh, "JHU-ISI gesture and skill assessment working set (JIGSAWS): A surgical activity dataset for human motion modeling," in *MICCAI Workshop: M2CAI*, vol. 3, p. 3, 2014.

[8] N. Madapana, M. M. Rahman, N. Sanchez-Tamayo, M. V. Balakuntala, G. Gonzalez, J. P. Bindu, L. N. V. Venkatesh, X. Zhang, J. B. Noguera, T. Low, R. Voyles, Y. Xue, and J. Wachs, "DESK: A Robotic Activity Dataset for Dexterous Surgical Skills Transfer to Medical Robots," *arXiv:1903.00959 [cs]*, Mar. 2019. arXiv: 1903.00959.

[9] R. DiPietro, C. Lea, A. Malpani, N. Ahmidi, S. S. Vedula, G. I. Lee, M. R. Lee, and G. D. Hager, "Recognizing Surgical Activities with Recurrent Neural Networks," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016* (S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, eds.), Lecture Notes in Computer Science, pp. 551–558, Springer International Publishing, 2016.

[10] N. Makondo, B. Rosman, and O. Hasegawa, "Accelerating model learning with inter-robot knowledge transfer," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2417–2424, May 2018.

[11] G. P. Moustris, S. C. Hiridis, K. M. Deliparaschos, and K. M. Konstantinidis, "Evolution of autonomous and semi-autonomous robotic surgical systems: a review of the literature," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 7, no. 4, pp. 375–392, 2011.

[12] C. E. Reiley, E. Plaku, and G. D. Hager, "Motion generation of robotic surgical tasks: Learning from expert demonstrations," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pp. 967–970, Aug. 2010.

[13] S. Krishnan, A. Garg, S. Patil, C. Lea, G. Hager, P. Abbeel, and K. Goldberg, "Unsupervised surgical task segmentation with milestone learning," in *Proc. Intl Symp. on Robotics Research (ISRR)*, 2015.

[14] 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.

[15] N. Padoy and G. D. Hager, "Human-machine collaborative surgery using learned models," in *2011 IEEE International Conference on Robotics and Automation*, pp. 5285–5292, IEEE, 2011.

[16] C. E. Reiley and G. D. Hager, "Task versus Subtask Surgical Skill Evaluation of Robotic Minimally Invasive Surgery," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009* (G.-Z. Yang, D. Hawkes, D. Rueckert, A. Noble, and C. Taylor, eds.), Lecture Notes in Computer Science, pp. 435–442, Springer Berlin Heidelberg, 2009.

[17] L. Tao, E. Elhamifar, S. Khudanpur, G. D. Hager, and R. Vidal, "Sparse Hidden Markov Models for Surgical Gesture Classification and Skill Evaluation," in *Information Processing in Computer-Assisted Interventions* (P. Abolmaesumi, L. Joskowicz, N. Navab, and P. Jannin, eds.), Lecture Notes in Computer Science, pp. 167–177, Springer Berlin Heidelberg, 2012.

[18] L. Tao, L. Zappella, G. D. Hager, and R. Vidal, "Surgical Gesture Segmentation and Recognition," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013* (K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, eds.), Lecture Notes in Computer Science, pp. 339–346, Springer Berlin Heidelberg, 2013.

[19] N. Ahmidi, L. Tao, S. Sefati, Y. Gao, C. Lea, B. B. Haro, L. Zappella, S. Khudanpur, R. Vidal, and G. D. Hager, "A Dataset and Benchmarks for Segmentation and Recognition of Gestures in Robotic Surgery," *IEEE Transactions on Biomedical Engineering*, vol. 64, pp. 2025–2041, Sept. 2017.

[20] C. Lea, G. D. Hager, and R. Vidal, "An Improved Model for Segmentation and Recognition of Fine-Grained Activities with Application to Surgical Training Tasks," in *2015 IEEE Winter Conference on Applications of Computer Vision*, pp. 1123–1129, Jan. 2015.

[21] C. Lea, R. Vidal, and G. D. Hager, "Learning convolutional action primitives for fine-grained action recognition," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1642–1649, May 2016.

[22] B. Béjar Haro, L. Zappella, and R. Vidal, "Surgical Gesture Classification from Video Data," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012* (N. Ayache, H. Delingette, P. Golland, and K. Mori, eds.), Lecture Notes in Computer Science, pp. 34–41, Springer Berlin Heidelberg, 2012.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.