Video-based Motion Expertise Analysis in Simulation-based Surgical Training Using Hierarchical Dirichlet Process Hidden Markov Model

Qiang Zhang, Baoxin Li
Computer Science and Engineering, Arizona State University
Tempe, AZ 85287-8809
qzhang53,baoxin.li@asu.edu

ABSTRACT

In simulation-based surgical training, a key task is to rate the performance of the operator, which is done currently by senior surgeons. This is a costly practice and objectively quantifiable assessment metrics are often missing. Researchers have been working towards building automated systems to achieve computational understanding of surgical skills, largely through analysis of motion data captured by video or other sensors. In this paper, we extend the Hierarchical Dirichlet Process Hidden Markov Model (HDP-HMM) for this purpose. We start with detecting spatial temporal interest points from the video capturing the tool motion of an operator, and then generate visual words from the descriptors of those interest points. For each frame, we construct a histogram with the associated interest points, i.e. the "bag of words", and then every video is represented by a sequence of those histograms. For sequences of each motion expertise level, we infer an HDP-HMM model. Finally, the classification of the motion expertise level for a testing sequence is based on choosing a model that maximizes the likelihood of the given sequence. Compared with the other action recognition algorithms, such as kernel SVM, our method leads to a better result. Further, the proposed approach also provides some important cues on the patterns of motion for each expertise level.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Motions*; I.5.2 [**Pattern Recognition**]: Design Methodology—*Classifier design and evaluation*

General Terms

Algorithm

MMAR'11, November 29, 2011, Scottsdale, Arizona, USA.

Keywords

Video analysis, motion expertise, surgery simulation, HDP-HMM, Dirichlet

1. INTRODUCTION

Understanding human motion is an important task in many fields such as sports, rehabilitation, computer animation and dance. One key problem is the analysis of motion expertise. In domains such as dance, sports and surgery, motion of experts differs considerably from that of novices. Based on this observation, several video-based motion analysis systems for sports analysis has been built. However those systems requires huge work on planning, calibration, and customization, making it difficult to extend them to other domains such as surgery. In recent years, surgical simulation has emerged at the forefront of new technologies for improving the education and training of surgical residents. The motivation of this paper is to develop computational algorithms that support the development of an intuitive and simple-to-use video-based system in the domain of simulation-based surgical training for motion expertise analysis.

The traditional process of training resident surgeons has been primarily based on interactive and direct instruction of supervising surgeons. As surgical motions have become increasingly complex and the surgeons have moved from open surgery to laparoscopic surgery to now robotic surgery, surgery education and training has become even more challenging. And thus conventional training purely relying on a senior surgeon's instruction is not only typically very slow, but also costly and hard to generalize since objectively quantifiable assessment metrics are often missing. Thus it is very important to find an efficient and effective computational approaches for the inference of surgical skills, where one challenge is to automatically rate the expertise level of a resident surgeon ([3]) from raw data captured during surgical training. This has currently gained added initiatives as the American college of surgeons seeks to develop a national skills curriculum (www.acs.org).

Objective evaluation of surgical skills has been a topic of research for many years ([4][6][9]). According to [1], there is high correlation between the expertise level and the motion parameters observed, such as duration, number of movements and length of path. This provides the theoretical foundation for building automated systems for objective evaluation of surgical skills using the collected features. A technique proposed in [3] called task deconstruction was implemented in a recent system reported in [8], where Markov

Models were used for decomposing a task (such as suturing) into basic gestures, and then the expertise level of the complex gesture could be analyzed. While this study offers an intriguing approach to expertise analysis, it requires an expert surgeon to provide specifications for building the topology of the model; hence it cannot be easily generalized to new procedures.

Most of the current systems require special devices, such as data gloves, which typically not only require some level of modification to the current training platform, but also may interfere with the performance of the subjects. Therefore, with the cameras strategically positioned without interfering the subject,a video-based system is preferred, in which the movement of the subject and/or the operating tools are captured and analyzed. While this problem is related to video-based action recognition, which has been under research for many years (some good surveys can be found in [11][7]), there are many new challenges. For example, most action recognition methods are designed for distinguishing different body movement patterns, whose visual motion is in general apparently different. In contrast, in the above problem of expertise level evaluation from surgical videos, the subjects are required to perform the same task and the difference of the visual motion due to their proficiency is often very subtle. Furthermore, even for subjects in the same proficiency level, they may show very different movements, due to the variation of personal habit. Thus typical action recognition methods do not directly apply for our task.

In this paper, we explore the above problem in the context of simulation-based laparoscopic surgery. The contribution of the paper comes with the following two aspects. First, we extend the HDP-HMM algorithm to learn discriminating models from the "bag of words" observations extracted from the raw videos; Second, in addition to assigning a label to the entire video (as done in a typical action classification method), the application of the extended HDP-HMM algorithm on motion expertise evaluation also provides many useful cues on the motion patterns of different expertise levels, by modeling the transitions among the features, which can lead to more useful analysis such as real-time feedback for improvement of a subject's action.

In Section 2, we introduce the feature, the HDP-HMM model and present our own approach. Then in Section 3, experimental results and analysis with real videos are reported. Section 4 concludes the paper with a brief discussion on future work.

2. PROPOSED APPROACH

2.1 Feature Extraction

"Bag of words", which represents each image (or video) as a histogram of visual words, is commonly used in object recognition (or action recognition) and promising results have been reported. In this paper, we adopt a similar idea to extract the features as follows. First, we use Laptev's space-time interest points toolbox [5] to detect interest points (with Histogram of Gradients (HoG) as the feature) from the video, where for a typical video, we obtain $10,000$ to $30,000$ interest points. Then we use PCA to reduce the dimension of the descriptors and use K-means to generate the visual words. For each frame, we find the interest points whose centers are on this frame and build a histogram of visual words with these interest points for this frame. Finally, each video is represented as a sequence of the histograms, which is different from a typical approach of action recognition with "bag of words". The proposed approach will model the transitions among these histograms and assign a label according to the transition.

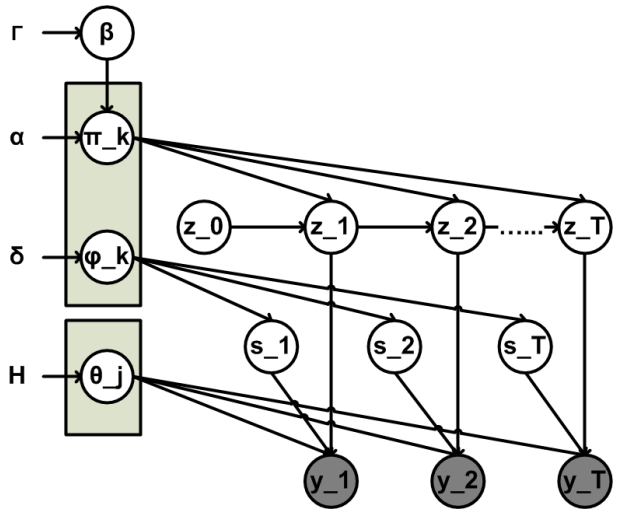


Figure 1: The graphic model for HDP-HMM, where the shaded circles are observed variables, unshaded circles are hidden variable and the un-circled are parameters.

2.2 HDP-HMM

Hierarchical Dirichlet process with hidden Markov model (HDP-HMM) was first proposed in [10] and applied in speaker diarization in [2]. In HDP-HMM, the transition between the states is sampled from a base distribution with a concentration parameter under Dirichlet Process. Fig. 1 shows the graphic model, where the shaded circles ($y_{1:T}$) are observed variables, unshaded circles (π_k, ψ_k, θ_k, $z_{1:T}$ and $s_{1:T}$) are hidden variable and the un-circled (H, β, α and δ) are parameters. These are further explained below:

1. $\beta \sim dp(\gamma)$: the global (average) transition probability;
2. $\pi_k \sim dp(\alpha, \beta)$: the tranistion probability for state k $k = 1, 2, ...$;
3. $\psi_k \sim dp(\sigma)$: the distribution of mixture components for state k $k = 1, 2, ...$;
4. $\theta_k \sim H$: the mixture components for $k = 1, 2, ...$;
5. $z_{t+1}|z_t \sim \pi_{z_t}$: the transition between states;
6. $s_t|z_t \sim \psi_{z_t}$: the selection of mixture components;
7. $y_t|z_t, s_t, \{\theta_j\}_{j=1}^{\infty} \sim F(\theta_{z_t,s_t})$: the emission at time t, e.g. $p(y_t|z_t, s_t, \{\theta_j\}_{j=1}^{\infty}) = N(y_t; \mu_{z_t,s_t}, \Sigma^2_{z_t,s_t})$ for the Gaussian observations, where $\theta_j = \{\mu_j, \Sigma_j^2\}$.

where $y_{1:T}$ denotes the sequence $y_1, y_2, ..., y_T$.

Compared with Hidden Markov Model (HMM), HDP-HMM has some attractive features. First, HDP-HMM is a Bayesian non-parametric approach: by defining a prior distribution over the transition probability matrices in a

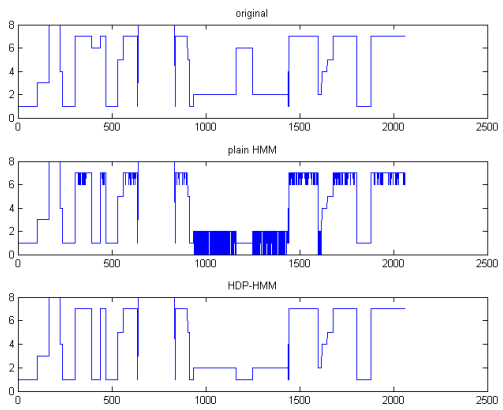


Figure 2: Simulation results demonstrating the recovery of the state sequence by HDP-HMM and plain HMM. The upper one is the ground truth, the middle one for the result obtained by HMM and the bottom one for HDP-HMM. The data was generated with 10 states and a single Gaussian observation model with 15−dimensional observation. The model was inferred from 100 sequences, each around 2000 frames long. Note, some states are not observed in this sequence. The HDP-HMM recovers the state sequence more accurate, eliminating many excessive switches in the state sequence.

countably infinite space, the HDP-HMM can have countably infinite number of states. Thus, unlike HMM, there is no need to fix the number of the states or the number of mixture models. In addition, HDP-HMM (or sticky HDP-HMM) uses a sticky parameter ρ to enforce the self transition, which can eliminate the fast switches in the state sequence recovered by HMM. We use a simulation to compare the recovered state sequences by HDP-HMM and plain HMM respectively (Fig. 2). Finally, HDP-HMM can be embedded in an on-line framework, which is very useful for a large data set. Given the existing data, we first infer the parameters and the variables, then update the priors for the parameters. These priors can be again used for future data.

Fox et al. proposed a modified forward-backward algorithm for the inference of the model: the state sequence $z_{1:T}$ is first inferred given the observation $y_{1:T}$; then the transition probability matrix $\{\pi_k\}$, the mixture components $\{\theta_k\}$ and other variables are learned; and finally the parameters and their priors are updated. The details of the inference algorithm can be found in [2].

2.3 HDP-HMM with Bag of Words

While the implementation of HDP-HMM by Fox et al. can infer the model with Gaussian mixtures, auto-regression and simple linear dynamic system observation model, we still need to extend it to work with "bag of words" observation model, which is widely used for video analysis task. In "bag of words", each frame is represented by a histogram of the visual words, where each bin indicates the occurrence of this visual word in this frame. Each histogram is drawn from a mixture component θ with $\sum_{i=1}^{N}\theta(i)=1$, where N is the number of visual words. Each mixture component is associated with one state. We can assume that the occurrence of a visual words in certain frame is independent of the other visual words, which indicates that the histogram is under categorical distribution. Thus the likelihood of state z given the observation y (histogram of one frame here) is

$$p(z|y)=\prod_{i}^{N}\theta_z^{y(i)}(i) \tag{1}$$

where θ_z is the mixture component associated with state z.

With Eqn. 1, we can infer the state sequence and transition probability matrix from the observation sequence by HMM inference methods, e.g. the backward-forward algorithm. By writing $\bar{Y}=\{y_t|z_t=z\ \forall t\}$ as the observations associated with state z among all the videos and assuming each of those observations are conditionally independent on the others given θ_z, i.e. $p(\bar{Y}|\theta_z)=\prod_{y\in\bar{Y}}p(y|\theta_z)$, θ_z can be updated by maximizing its posterior:

$$p(\theta_z|\bar{Y},H)=\frac{p(\bar{Y}|\theta_z)p(\theta_z|H)p(H)}{\int_{\theta_z}p(\bar{Y}|\theta_z)p(\theta_z|H)p(H)}\propto p(\theta_z|H)\prod_{y\in\bar{Y}}p(y|\theta_z) \tag{2}$$

where we have marginalized out the other parameters, variables and also made the assumption that $\bar{Y}$ is conditionally independent of H given θ_z, i.e.
$p(\theta_z,\bar{Y},H)=p(\bar{Y}|\theta_z)p(\theta_z|H)p(H)$.

If $p(\theta|H)$ is a Dirichlet distribution, i.e. $p(\theta|H)\sim dir(H)$, we can obtain a closed form for Eqn. 2 by using the conjugate property between categorical distribution and Dirichlet distribution:

$$p(\theta_z|\bar{Y},H)=dir(H(1)+m(1),H(2)+m(2),...,H(N)+m(N)) \tag{3}$$

where dir is the Dirichlet distribution and $m(i)=\sum_{y\in\bar{Y}}y(i)$ is the total count of the occurrences of visual word i in all the observations associated with component θ_z.

2.4 Classification with Inferred Model

To apply the HDP-HMM for classification task, for sequences of each class, we learn a model with the inference algorithm of HDP-HMM. Given a new testing sequence, we apply Viterbi algorithm to find an optimal path given each learned model. The classification is based on choosing the model that gives the maximal likelihood of the optimal state sequence:

$$p(y_{1:T}|z_{1:T},\{\pi_k\}_{k=1}^{K_z},\{\theta_j\}_{j=1}^{K_s})= \tag{4}$$
$$\pi_{z_1}p(y_1|\theta_{z_1})\prod_{t=2}^{T}\pi_{z_{t-1}}(z_t)p(y_t|\theta_{z_t})$$

where $z_{1:T}$ is the optimal state sequence obtained by Viterbi algorithm, K_z for the inferred number of states and K_s for the inferred number of mixture components.

2.5 Comparison between HDP-HMM with Bag of Words and HDP-HMM with Mixture Gaussian Observation Model

Compared with HDP-HMM with mixture Gaussian observation model, the proposed HDP-HMM with bag of words has several benefits. First, HDP-HMM with bag of words is a more concise model and requires much less variables. For each state, HDP-HMM with bag of words only need to maintain a single vector, i.e. $\theta_z\in\mathbb{R}^{d\times 1}$, for the mixture

component, where d is the dimension of observation. In the other hand, for HPD-HMM with mixture Gaussian observation mode, we need to maintain the mean ($\in \mathbb{R}^{d\times 1}$) and variance matrix ($\in \mathbb{R}^{d\times d}$) for each state. As a result, the inference of HDP-HMM with bag of words is more stable than HDP-HMM with mixture Gaussian observation, when only limited training data is available, like in the application of this study.

Second, one difficulty of applying HDP-HMM is how to select a good priors for the parameter, e.g. the prior distribution H for the mixture components. For HDP-HMM with mixture Gaussian observation model, H includes the prior distributions for the mean and the covariance matrix of the mixture component, which are difficult to obtain. For HDP-HMM with bag of words, the prior H can be simply set to Dirichlet distribution with a homogeneous vector, which means all visual words have equal probability at initialization. Thus, it is easier to set the priors of the parameters for the HDP-HMM with bag of words than for the HDP-HMM with mixture Gaussian observation model.

Third, the HDP-HMM with bag of words is more robust in face of data noise. The estimation of Gaussian mixture is sensitive to the outliers/noise in the data. However, as we use bag of words, the outliers/noise will only affect a small portion of the histogram. Thus, the inference of HDP-HMM with bag of words is less sensitive to the noise/outliers.

3. EXPERIMENT

Videos captured from resident surgeons in two local hospitals were used in our experiments to evaluate the proposed approach in classifying the expertise level. In the following, we first briefly describe the data used in our experiments and then present some analysis results.

We used 28 videos for the "peg transfer" operation by 28 different subjects with 2 proficiency levels: 15 of the resident surgeons are deemed as experts who are very skilled with the task while the other 13 are considered as novices who are yet to gain better skills with the task. Peg transfer is one of the standard training tasks a resident surgeon needs to perform and pass. This task requires the subjects to lift (i.e., "Pick") six objects (one by one) with a grasper by the non-dominant hand, transfer the object midair to the dominant hand, and then place (i.e., "Drop") the object on a peg on the other side of the board. Once all six objects are transferred, the process is reversed, and the objects are to be transferred back to the original side of the board. The timing for this task starts when the subjects grasped the first object and ends upon the release of the last peg.

The length of the captured videos vary from 1500 frames to 5500 frames. We split the videos into 900-frame clips, with 50% overlapping, which results in 201 clips with 69 clips deemed at expert level and the other 132 at novice level. We have manually assigned the labels for the videos and the clips based on guidelines and information provided by the collaborating hospitals. Fig. 3 illustrates one sample frame from a video in the experiment. Tab. 1 provides some details of the data we used in the experiment.

For interest point detection, we use 3-level pyramids (after resizing frame resolution to 360×240) and HoG as the feature. We have $10,000$ to $30,000$ interest points for each video. We then use PCA to reduce the dimension of the descriptor to 15 and use K-means to generate 400 visual words. For the frames with no interest points associated, we eliminate them, since they have no affect on the inference of the model. Finally, each clip is represented as a sequence of histograms, with varying length. For HDP-HMM inference, we set the cutting level for the number of states to 10 and the priors for the parameters are $H = 2$, $\alpha \sim \Gamma(1, 0.01)$, $\gamma \sim \Gamma(6, 1)$ and $\rho \sim \beta(0, 1)$.

Figure 3: A sample frame from one of the videos used in the experiment. The peg transfer task requires the subjects to lift (i.e. "Pick") six objects (one by one) with a grasper by the non-dominant hand, transfer the object midair to the dominant hand, and then place (i.e. "Drop") the object on a peg on the other side of the board. Once all six objects are transferred, the process is reversed.

Table 1: The details of the video we used in the experiment. The resolution is 720×480 and 25 FPS for frame rate. We down sample all the video to 360×240 in the experiment.

Level	# videos	# Frames	# clips	Total # clips
Expertise	15	1500-4000	3-7	69
Novice	13	3700-5500	7-12	132

The proposed approach is compared to HDP-HMM with mixture Gaussian as observation model and kernel SVM. For kernel SVM, we build a histogram for each clip by using all the interest points in this clip. Then we use RBF-χ^2 kernel:

$$\kappa(x, y) = e^{-\frac{1}{2\sigma}\times\sum_i \frac{(x_i-y_i)^2}{x_i+y_i}} \quad (5)$$

where σ is the selected to be the average distance. In addition, to alleviate the affect of choosing particular σ, we also use AdaBoost to combine multiple kernels with different σ: $[\frac{1}{4}\sigma_0, \frac{1}{2}\sigma_0, \frac{3}{4}\sigma_0, 1\sigma_0, 2\sigma_0, 3\sigma_0, 4\sigma_0]$, where σ_0 is the average distance as used in kernel SVM.

For HDP-HMM with mixture Gaussian as observation model, we use optical flow as feature. For each frame, we compute the optical flow at 30×45 regular grid, which results in a $\mathbb{R}^{1350\times1}$ feature vector. To reduce the dimension, we further apply PCA and obtain $\mathbb{R}^{30\times1}$ feature vector. Then we use K means ($K = 50$) to analysis the mean and variance (Σ) of the data, also get the estimation of truncation level of the number of states. The number of states is truncated at $K_z = 10$, while the number of mixture components is truncated at $K_s = 50$. The mixture components are drawn from Normal-Inverse-Wishart prior (conjugate distribution to Gaussian distribution): mean is under normal distribution with zero mean and Σ as variance; variance is under inverse Wishart distribution with $0.2 \times I_d$ as the mean and $d + 2$ degree of freedom, where d is the feature dimension,

Table 2: The experiment result of the HDP-HMM and the kernel SVM, where the result of the proposed approach is in Row 2, Row 3 for the result of HDP-HMM with mixture Gaussian model, Row 4 for the result of kernel SVM and Row 5 for the result of Kernel SVM with AdaBoost. All the results are the average of 3 rounds of experiments

Clip			Video		
Expert	Novice	Total	Expert	Novice	Total
70.31%	95.77%	86.27%	85.71%	92.86%	89.29%
76.32%	56.06%	65.38%	66.67%	66.67%	66.67%
43.94%	94.36%	78.37%	50.00%	100.00%	73.08%
53.24%	93.57%	81.19%	64.29%	100.00%	80.77%

i.e. 30. The other priors are the same as HDP-HMM with bag of words.

We use "leave one subject out" scheme, i.e. for each round of experiment, we use the clips of one video from each level for testing and the remaining for training. This can remove the bias introduced by the over-lapping during cutting the videos into clips. We perform 3 rounds of experiment and compute the average, where in each round, the clips of each subject would be used for testing at least one time.

The accuracy contains two parts: the accuracy on clip level and the accuracy on video level, where the label of the testing video is decided by the majority voting with its clips' labels, which is shown in Tab. 2. From this table, we can find that the proposed method outperforms kernel SVM and kernel SVM with AdaBoost, especially on the clips and videos at expert level. Besides, the HDP-HMM with mixture Gaussian observation model gets a very low performance. In fact, the performance of HDP-HMM with mixture Gaussian observation model on clip level is around the random guess ($\frac{132}{201} = 65.67\%$). As discussed in Sec. 2.5, it may due to the reason that, the inference of HPD-HMM with mixture Gaussian observation model is not stable, given the limited training data.

Along with better performance, the proposed approach also gives us some meaningful insights of the action patterns of operators at expertise level and novice level. Fig. 4 shows the state transition probability for an expert (a) and a novice (b). From this figure, we can find that, the state sequence for an expert tends to have more switches between states than that for a novice (i.e. a lower probability for self transition). Fig. 5 shows the examples of the inferred state sequences for clips at the expert level (top rows) and the novice level (bottom rows). This figure verifies our observation from Fig. 4 that state sequence for expert has more transitions among the states than that of expert, which may be explained as, the operator at the novice level needs more time to finish certain movements, which corresponds to the "flat" state sequence in the bottom row of Fig. 5; on the other hand, the operator at the expert level can finish movement much more faster, which results in the fast switches among the states.

4. CONCLUSION AND FUTURE WORK

In this paper, we extended the HDP-HMM approach to "bag of words" observation model and applied it on motion expertise evaluation using videos captured in simulation-based surgical training. By modeling the transitions in the sequence, the proposed method not only out-performs the kernel-SVM, which simplifies the video to a single histogram, but also provides some useful cues on the motion pattern of different expertise level. These patterns could help in building an future on-line system, which can give live feedback to the operator, thus facilitating the training of resident surgeons. In addition, the proposed method is not specialized for the specified action type, and thus it can be used to analysis the other action types. In the future, we will work on how to combine different modality of feature to get a better performance, e.g. how to utilize the videos capturing the movement of the hands of the operators.

5. ACKNOWLEDGMENTS

The authors were partially supported during this work by a grant from NSF (Award # 0904778), which is greatly appreciated.

6. REFERENCES

[1] R. Aggarwal, T. Grantcharov, K. Moorthy, and T. Milland. An Evaluation of the Feasibility , Validity , and Reliability of Laparoscopic Skills Assessment in the Operating Room. *Annals of Surgery*, 245(6):992–999, 2007.

[2] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. A sticky hdp-hmm with application to speaker diarization. *arxiv*, 2010.

[3] A. G. Gallagher, E. M. Ritter, H. Champion, G. Higgins, M. P. Fried, G. Moses, C. D. Smith, and R. M. Satava. Virtual Reality Simulation for the Operating Room. Proficiency-Based Training as a Paradigm Shift in Surgical Skills Training. *Annals of Surgery*, vol:241pp364–372, 2005.

[4] A. N. Healey, U. S, and C. A. Vincent. Developing observational measures of performance in surgical teams. *Qual. Saf. Health Care*, vol:13pp33–40, 2004.

[5] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2):107–123, 2005.

[6] S. Mayes, J. Deka, K. Kahol, M. Smith, J. Mattox, and A. Woodwards. Evaluation Of Cognitive And Psychomotor Skills Of Surgical Residents at Various Stages in Residency. *5th Annual Meeting of American College of Obstetricians and Gynecologists*, 2007.

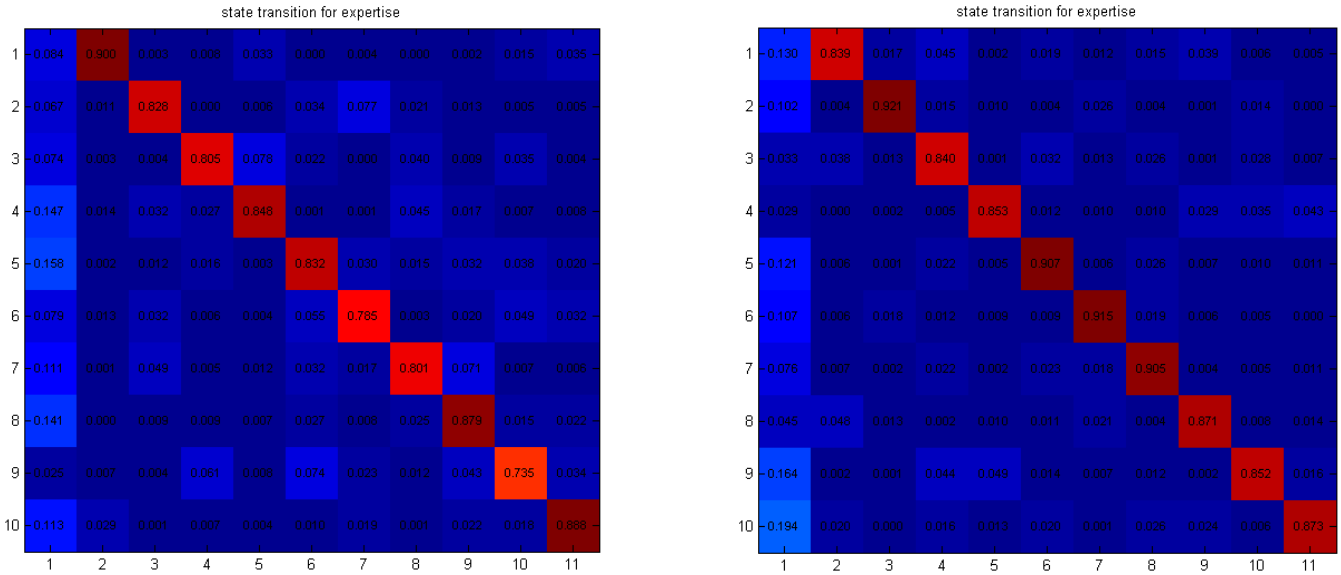
[7] S. Mitra and T. Acharya. Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 37(3):311–324, 2007.

[8] J. Rosen, J. D. Brown, L. Chang, M. N. Sinanan, and B. Hannaford. Generalized approach for modeling minimally invasive surgery as a stochastic process using a discrete markov model. *IEEE Transactions on Biomedical Engineering*, 53(3):399–413, 2006.

[9] R. M. Satava, A. G. Gallagher, and C. A. Pellegrini. Surgical competence and surgical proficiency: definitions, taxonomy, and metrics. *Journal of the American College of Surgeons*, 196(6):933–937, 2003.

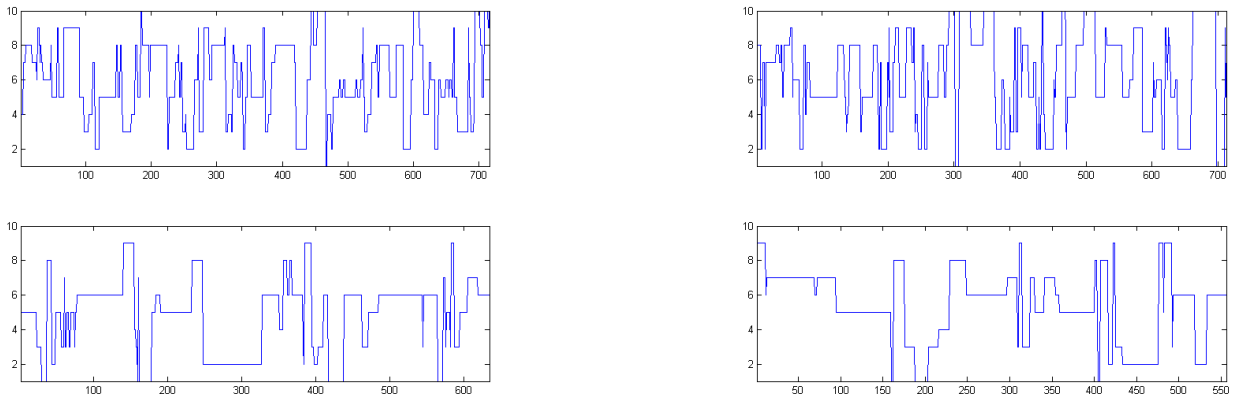
[10] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical*, pages 1–34, 2006.

[11] J. J. Wang and S. Singh. Video analysis of human dynamics–a survey. *Real-Time Imaging*, 9(5):321–346, 2003.



(a) (b)

Figure 4: The state transition probability matrix for expert (a) and novice (b), where the Col. 1 of each plot is the probability for initial state. For Col. 2 to Col. 11, the element in Row i Col. j means the probability of transiting from state i to state j. This figure is better viewed in color. From this figure, the elements on the diagonal of (a) are generally smaller than those on the diagonal of (b).



(a) (b)

Figure 5: This figures shows the inferred state sequence for 4 clips, where the top rows are for clips labeled as expert and bottom rows for novice. The x-axis is frame number and the y-axis the state number. Please note, we remove the frame which has no feature points associated, so the number of frames in each clip is usually smaller than 900. This figure shows that the state sequence from videos of expert (top row) has obviously more transitions than that of novice (bottom row).