INSTRUCTIVE VIDEO RETRIEVAL FOR SURGICAL SKILL COACHING USING ATTRIBUTE LEARNING

Lin Chen, Qiang Zhang, Peng Zhang, Baoxin Li

Computer Science and Engineering, Arizona State University
{lin.chen.6, qzhang53, pzhang41, baoxin.li}@asu.edu

ABSTRACT

Video-based coaching systems have seen increasing adoption in various applications including dance, sports, and surgery training. Most existing systems are either passive (for data capture only) or barely active (with limited automated feedback to a trainee). In this paper, we present a video-based skill coaching system for simulation-based surgical training by exploring a newly proposed problem of instructive video retrieval. By introducing attribute learning into video for high-level skill understanding, we aim at providing automated feedback and providing an instructive video, to which the trainees can refer for performance improvement. This is achieved by ensuring the feedback is weakness-specific, skill-superior and content-similar. A suite of techniques was integrated to build the coaching system with these features. In particular, algorithms were developed for action segmentation, video attribute learning, and attribute-based video retrieval. Experiments with realistic surgical videos demonstrate the feasibility of the proposed method and suggest areas for further improvement.

Index Terms— Instructive Video Retrieval, Attribute Learning, Coaching System

1. INTRODUCTION

Video-based coaching systems aim at helping people to improve their skills through capturing their performance via video recordings that allow either on-line or off-line analysis. Applications of such systems include dance [1][2], sports [3], and machine-operation training [4], etc. Traditionally, the analysis is performed only by humans (e.g., coaches and trainers). Recent years have witnessed increasing interests in developing automated systems for doing such analysis for improved training, where the key task is vision-based motion skill understanding since the coaching task often boils down to providing corrective feedback to a trainee regarding his/her movements.

Most existing methods may provide one of the following two types of feedback. The first type is an overall assessment with either a numeric rating [2] or a skill level [5]. While being useful for skill examinations, such type of feedback provides little suggestion to a trainee as to how to improve. Further, many methods [1][2][5] are based on comparison using some sort of standard action series or models. This limits the applicability of such methods to complex tasks where defining a standard action or model is impractical due to the existence of a wide range of valid/perfect solutions. The second type of feedback is some statistics computed from a user's training sessions, such as total execution time, movement counts, motion smoothness, etc. Although such statistics are more informative, they do not readily lead to corrective actions that the trainees may take to improve their performance.

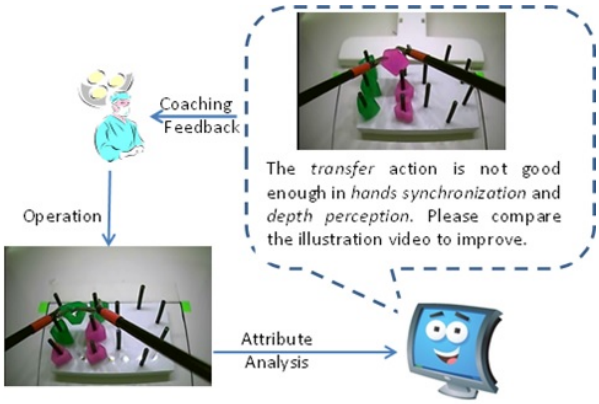
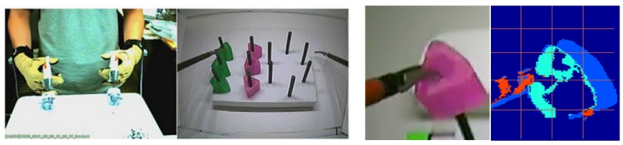


Fig. 1. Illustration of the proposed skill coaching system that retrieves an illustrative video as feedback while providing specific and expressive verbal suggestions.

In this paper, we present a video retrieval system (illustrated in Figure 1) for skill coaching in simulation-based surgical training, which we defined as **instructive video retrieval**. We aim at providing automated video and verbal feedback that has the following three features: (1) **specificity**: the feedback should focus on a trainee's skill weakness; (2) **superiority**: the retrieved illustrative video should represent a better skill than the trainee; (3) **similarity**: the retrieved illustrative video should have a similar operation context to the trainee's video. Note that although the focus of the paper is on the specific application of simulation-based surgical training, the above features are deemed as critical to effective skill coaching in general [6], and thus the proposed method can be extended to other video-based skill coaching applications by



(a) (b)

Fig. 2. The illustration of (a) the FLS system; (b) object-motion distribution for action recognition.

integrating corresponding domain knowledge.

Different from traditional video retrieval such as nearly-duplicated video retrieval [7] or concept retrieval [8], which are purely based on video content, the instructive video retrieval requires both low-level content analysis and high-level semantic skill understanding. To our knowledge, this is still a new research effort with little prior art. In this work, we introduce semantic attributes in video to bridge the gap between inherent vagueness and the subjectivity of "instructive".

The technical contribution of this work is threefold. First, we propose a new video retrieval problem defined as "instructive video retrieval" and a corresponding effective algorithm to solve this problem. This new problem has a wide variety of applications and more efforts are worth investing. Second, we extend image attribute learning into the video domain for skill evaluation, which is useful to bridge the gap between the low-level motion measurements and high-level skill understanding. Third, we develop a vision-based skill coaching system for simulation-based skill training, which provides an automatic and efficient way for self skill improvement without costly human supervision.

This study is primarily based on surgical training on the Fundamentals of Laparoscopic Surgery (FLS) trainer box (www.flsprogram.org), a simulation-based platform that has been widely used in many hospitals for minimally-invasive surgery training. The system is essentially a box simulating the human body and a trainee is required to use tools going into the box through small holes to perform actions like lifting and transferring objects inside the box (Fig. 2(a) left). The trainee can see what is going on inside the box only through a monitor that displays a live video (Fig. 2(a) right) captured by an on-board camera. In the operation, a trainee is required to lift one of the six objects with a grasper in his non-dominant hand, transfer the object midair to his dominant hand, and then place the object on a peg on the other side of the board. Once all six objects have been transferred, the process is reversed from one side to the other.

2. PROPOSED METHOD

In this section, we present our proposed method for video-based skill coaching which performs three key tasks: (1) Decomposing a video clip into primitive action units; (2) Rating each action unit using semantic attributes; (3) Retrieving an illustrative video for instruction. Fig. 3 presents a flow chart of our system, outlining its major algorithmic components and their interactions.

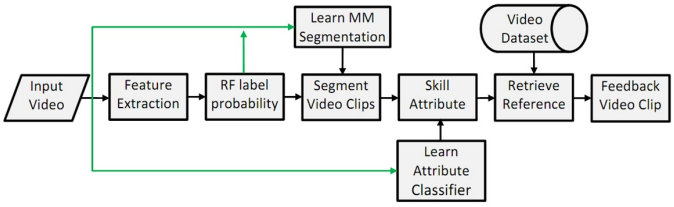


Fig. 3. An overview of the proposed approach. The green components are only used in the training stage.

2.1. Primitive Action Segmentation

We first segment the given video into clips where each clip only includes one primitive action. The FLS operation consists of 5 primitive actions [9] as building blocks of manipulative surgical activities: (**L**): lift an object from the peg, (**T**): transfer an object, (**P**): place an object on the peg, (**W**): move the grasper with an object, (**U**) move the object without an object. Since the videos we consider exhibit predictable motion patterns arising from the underlying actions of the human subject, we adopt the Hidden Markov model (MM) in the segmentation task. This allows us to incorporate domain knowledge into the transition probabilities, e.g., the Lift action is followed by itself or by Loaded Move with high probability. Following [10][4], we define each state as a primitive action. The task of segmentation is then to find the optimal state path for the given video.

Frame-level Feature Extraction The FLS box is a controlled environment including 4 types of objects: background, rubber cubes, pegs, and tools/graspers. Due to the noisy video clips, we designed a probability representation of the motion information which served as features for action segmentation. Specifically, we first use random forest (RF) [11] to obtain the label probability $p_l(x)$ that a pixel x belongs to the object label l. Then the tool orientation and tip region can be further detected by the spatial information based on the obtained probabilities. Since all surgical actions occur in the region around the grasper tip, the region is defined as the ROI region (Fig. 2(b) left) to filter out other irrelevant background. With comparison with the distribution of the background region, we estimate the probability of each pixel x being "moving" by frame differencing, which is denoted by $p_m(x)$. Then the joint distribute $p_l(x) \cdot p_m(x)$ represents the probability that the pixel x is the moving object l. This joint object-motion distribution suppresses the static clutter background in the ROI so that only interested motion information will be reserved. To further capture the spatial information, we further split the ROI into blocks, as shown in Fig. 2(b) right, and describe the object-motion distribution in each block by the Hu-invariant moment [12]. Finally the moment vectors in each block are cascaded into a frame-level descriptor for action recognition.

Random Forest as Observation Model After obtaining

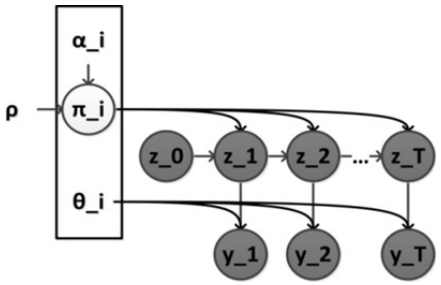


Fig. 4. The graphical model for Bayesian estimation of transition probability, where the symbols with circles are hidden variable to be estimated, the symbols within gray circle are observations and the symbols without circle are priors.

the frame-level descriptor, we further utilize RF for frame-level action recognition. Since RF is an ensemble classifier with a set of decision trees and the output is based on majority voting of the trees in the forest, the frame-level action distribution provides a good estimation of the observation model for the HMM states. Assuming that there are N trees in the forest and n_i decision trees assign primitive action label i to the input frame, we could view random forest chooses label i with probability n_i/N, which can be taken as the observation probability for State (primitive action) i.

Bayesian Estimation of Transition Probability The transition probability from State i to State j can be estimated based on a small set of data as the ratio the number of (expected) transitions from i to j over the total number of transitions. However, one potential issue of this method is that, in video segmentation we have limited training data in video segmentation. Furthermore, the number of transitions among different states (primitive action), is typically much less than the total number of frames of the video. This will result in a transition probability matrix which is dominated by diagonal elements. The resulting transition probability will degrade the benefit of using HMM for video segmentation, i.e., forcing desired transition pattern in the state path. Thus, we propose to use a Bayesian approach for estimating the transition probability, employing the Dirichlet distribution which enables us to combine the domain knowledge with the limited training data for transition probability estimation. The model is shown in Fig. 4.

Assuming $\alpha_i(\sum_j \alpha_i(j) = 1)$ is our domain knowledge for the transition probabilities from State i to all states, then we can draw the transition probability vector π_i as $\pi_i \sim dir(\rho\alpha_i)$ where dir is the Dirichlet distribution as a distribution over distribution, and ρ represents our confidence of the domain knowledge. Given the transition probability π_i, the count of transition from State i to all states follows a multinomial distribution:

$$n_i \sim multi(n_i|\pi_i) = \frac{(\sum_j x_i(j))!}{\prod_j n_i(j)!} \prod_j \pi_i(j)^{n_i(j)}. \quad (1)$$

Because the Dirichlet distribution and multinomial distribution are a conjugate pair, the posterior probability of transition probability is just combining the count of transition among state and domain knowledge (prior) as $\pi \sim dir(n_i + \rho\alpha_i)$. When there are not enough training data, i.e., $\sum_i n_i(j) \ll \rho$, π_i would be dominated by α_i, i.e., our domain knowledge; as more training data become available, π_i would approximate to the counting of transitions in the data.

Table 2. Motion measurements.

Motion	Description	Definition
Instrument	The motion of grasper tip.	$v(t)$
Instrument target	Relative motion between grasper tip and its operation target.	$\hat{v}(t)$ $v_r(t)$
Object	The motion area of objects in ROI.	$A(t)$
ROI	The optical flow field in ROI.	$m(x,t)$

2.2. Attribute Learning for Action Rating

A fundamental challenge in instructive video retrieval is to map computable visual features to semantic concepts that are meaningful to a trainee. Recognizing the practical difficulty of lacking sufficient amount of exactly labeled data for learning an explicit mapping, we introduce the concept of image attribute into video domain to evaluate the underlying skill of an action clip based on semantic attributes designed using domain knowledge. Following [13][14][15], we define 6 attributes to measure the skill level of the trainee's operation, which are listed in Table 1. With these semantic attributes, the system will be able to expressively inform a trainee what is the weakness in the operation, since the defined attributes are all semantic concepts used in existing human-expert-based coaching (and thus they are well understood).

Feature Representation To represent the defined attributes, new motion features are constructed for skill rating in 3 steps. First, a few types of motion measurements, which are summarized in Table 2, are calculated based on the previous object segmentation information (Sect. 2.1). Second, we extract motion signatures from each of the motion measurement, which are summarized in Table 3. The motion signatures are 1-dimensional temporal signals that further compact the motion information. Last, final motion feature are constructed from each motion vector and its motion signatures as follows. In the temporal domain, we divide a signature into equal temporal bins; in the Fourier domain, we also divide the

Table 3. Motion signatures. $y(t)/y(x)$ represents any motion vector in Table 2, e.g. $v(t)$, $v_r(t)$, $A(t)$, etc. $\bar{y}(t)$ is the smooth result of $y(t)$. m is the shorthand for field motion $m(x,t)$.

Name	Definition	Description
Velocity	$\lvert y(t)\rvert$	Instant velocity
Path	$\int_0^t \lvert y(x)\rvert dx$	Accumulated motion energy
Jitter	$\lvert y(t) - \bar{y}(t)\rvert$	Motion smoothness metric
CAV	$\int \frac{\lvert\langle \nabla \times m, m\rangle\rvert}{\lVert m\rVert_2} dx$	Curl angular velocity

Table 1. Action attributes for surgical skill assessment.

Attributes	Description
Time and motion (**T**)	How efficiently a trainee can operate without unnecessary moves.
Flow of operation (**F**)	How smoothly a trainee can operate without frequently stops.
Bimanual dexterity (**B**)	How well two hands can cooperate and work together.
Respect for issue (**R**)	How force is controlled in operation of objects as subjective evaluation of organ damage.
Instrument handling (**I**)	How well a trainee operates instruments without bad attempts and movements.
Depth perception (**D**)	How good a trainee's sense of depth to avoid failed operation on a wrong depth level.

frequency into equal bins. In each temporal and frequency bin, the maximal, minimal, and average values are cascaded into the final feature set.

Relative Attribute Skill Rating To rate the skills of each primitive action video clip, we adopt relative attribute learning [16] to calculate relative ranking of the clips with respect to the defined semantic attributes. Formally, for the k-th attribute, we are given a set of ordered pairs of clips $O_k = \{(i,j)\}$ and a set of un-ordered pairs $S_k = \{(i,j)\}$, where $(i,j) \in O_k$ means the video clip v_i has a better skill performance than the video clip v_j (i.e. $v_i \succ v_j$) in terms of the specified attribute, and $(i,j) \in S_k$ means v_i and v_j have similar skill performance (i.e. $v_i \sim v_j$). Then the constructed motion features can be fed into the following relative attribute framework to learn the model $\boldsymbol{w}_k$ for relative skill rating:

$$\begin{aligned} \min_{\boldsymbol{w}_k,\varepsilon,\gamma} \; & \frac{1}{2}\|\boldsymbol{w}_k^T\|_2^2 + C(\sum \varepsilon_{ij}^2 + \sum \gamma_{ij}^2) \\ s.t. \; & \boldsymbol{w}_k^T(\boldsymbol{x}_i - \boldsymbol{x}_j) \geq 1 - \varepsilon_{ij}, \forall (i,j) \in O_k; \\ & |\boldsymbol{w}_k^T(\boldsymbol{x}_i - \boldsymbol{x}_j)| \leq \gamma_{ij}, \forall (i,j) \in S_k; \\ & \varepsilon_i \geq 0, \gamma_{ij} \geq 0. \end{aligned} \tag{2}$$

where $\boldsymbol{x}_i$ is the feature vector extracted from the i-th video clip, C is the trade-off constant to balance maximal margin and pairwise attribute order constraints. The relative skill can be compared by the attribute value $\boldsymbol{w}_k^T \boldsymbol{x}_i$, which is used in the subsequent retrieval of illustrative videos.

2.3. Illustrative Action Clip Retrieval

With the previous processing, the system will retrieve an illustrative video clip from a constructed video repository and present it to a trainee as an instructive video.

Operation Weakness Detection We first need to figure out what is the operation weakness. However, a lowest attribute value does not always mean the most urgent attribute in need of improvement, especially when this attribute is "difficult" to most of the people. The weakest attribute should be the one that the trainee did poorly while most other people are significantly better. Thus, we use the average cumulative distribution of one user's training session to assess the performance strength of each attribute. Specifically, with K attributes, each clip v_i can be characterized by a K-dimensinoal vector $[a_{i,1}, \cdots, a_{i,K}]$, where $a_{i,k} = \boldsymbol{w}_k^T \boldsymbol{x}_i$ is the k-th attribute value of v_i based on its feature vector x_i. The attribute values of all clips (of the same action) in the data repository forms an $N \times K$ matrix A whose column vector a_k is the k-th attribute value of each clip. Similarly, from a user's training session, for the same action under consideration, we have another set of clips with attribute matrix $\hat{A}$ whose column vector $\hat{a}_k$ is the user's k-th attribute values in the training session. The performance strength of the k-th attribute s_k can be calculated by

$$s_k = \frac{1}{n} \sum_{i=1}^{n} P(\boldsymbol{a}_k \geq \hat{a}_{i,k})) \tag{3}$$

where n is the total number of video clips in one training session of one primitive action. Note that higher s_k means more users are doing better than the current operation, which means high importance the attribute need to improve.

Video Utility Evaluation We then need to figure out how much a video clip v_i is helpful with regard to a given training video, which we defined as the utility of v_i on the k-th attribute, denoted as $u_{i,k}$. We measure the utility by normalized distance between the performance strength of these two videos. Specifically, let s_k and $\hat{s}_k$ denotes the performance strength of the input training video and the potential instructive video clip v_i on the k-th attribute, the utility is calculated as

$$u_{i,k} = \frac{s_k - \hat{s}_k}{s_k} \tag{4}$$

After the calculation of these measurement, we select the best illustration video clip v_i^* from the data repository using the following criterion:

$$v_i^* = \arg\max_i \sum_{k=1}^{K} s_k \cdot u_{i,k} \tag{5}$$

The underlying idea of Function (5) is that a good feedback video should have high utility on important attributes.

With the above attribute analysis, we also provide a verbal description with regard to the 3 worst action attributes with an absolute importance above a threshold 0.4, which means that more than 60 percent of the pre-stored action clips are better in this attribute than the trainee. If all attribute importance values are lower than the threshold, we simply select the worst one. With the selected attributes, we retrieve the illustration video clips, inform the trainee about on which attributes he performed poor, and direct him to the illustration video.

It is worth noting that in recommending an illustration video, we defined concepts that are context dependent. That is, the importance and utility values of an attribute depends on the given data set. In practice, the data set could be a local database captured and updated frequently in a training center, or a fixed standard dataset, and thus the system allows adjustment of some parameters (e.g., the threshold 0.4) based on the nature of the database.

3. EXPERIMENTS

We tested our framework on a 64-bit computer with Intel Core i5-2500 CPU @ 3.30GHz and 8.00G RAM. Experiments have been performed using realistic training videos capturing the performance of resident surgeons (includes experts and novices) in a local hospital during their routine training. A typical testing video contains about 4500 frames with the frame rate of 30 FPS and the resolution of 480×720. For acceleration, we down-sampled the resolution and frame rate by 2. Experiments showed that our system takes on average around 242 seconds to process such a video.

For evaluating the proposed methods, we selected 10 representative videos/subjects from trainees of different skill levels. Each video is a full training session consisting of 12 Peg Transfer cycles which leads to 12 video clips for each primitive operation. We have a database of 240 clips for each of the 3 primitive operations, i.e. lift, transfer, and place, with the total number of clips being 720. The video can be downloaded from zhqiang.org/?p=207. We emphasize that, even the same subject does not perform the same action identically (and in fact the variability was observed to be very high), and thus the clip dataset is very diverse, providing a reasonable basis for evaluating our method. The exact frame-level labeling (which action each frame belongs to) was manually obtained as the ground truth. For each primitive action, we randomly select 150 pairs of video clips and then manually label them by examining all the attributes previously defined. This process manually determines which video in a given pair should have a better skill according to a given attribute.

3.1. Action Segmentation

As discussed in the previous section, we first calculated the frame-level action classification accuracies, then the classification scores (probabilities) are used as the observation into an HMM to get a final action recognition result, which is solved by the Viterbi algorithm. The confusion matrix of action recognition before and after employing HMM is shown in Table 4, which is calculated by leave-one-video-out cross validation. It can be seen that the frame-level recognition result is already high for some actions, which verifies the effectiveness of our proposed motion descriptors. The recognition accuracies after using the HMM are significantly improved, especially for actions L and P. The overall low accuracy for actions L and P is mainly due to the trainee's unsmooth operation that caused many unnecessary stops and moves, which are hard to distinguish from UM and LM. The overall segmentation accuracy of expert videos is 93.5% while the accuracy of novice videos is 80.3%. The results show that the proposed action segmentation method is able to deliver reasonable accuracy in face of some practical challenges.

Table 4. Confusion matrix of the action recognition accuracies. Each cell is the accuracy (%) with/without HMM.

	UM	L	LM	T	P
UM	87.6/88.0	0.2/0.2	0.6/0.8	11.5/10.3	0.8/0.8
L	21.9/36.1	43.4/28.5	21.8/15.8	13.0/13.3	0.0/6.3
LM	3.8/18.0	0.2/1.1	77.3/61.1	12.8/12.3	6.0/7.5
T	5.6/11.3	0.0/0.1	1.0/0.9	93.4/87.7	0.0/0.0
P	28.7/55.1	0.6/2.8	12.0/19.9	1.3/2.4	57.5/19.9

Table 5. Accuracy of attribute learning across primitive actions (%). Each row represents a different primitive action and each column represents a different attribute.

	T	F	B	R	I	D
L	92.7	95.1	97.2	92.5	97.4	87.5
T	90.0	97.9	82.9	N/A	N/A	92.5
P	83.0	89.5	88.2	95.2	95.8	89.8

3.2. Skill Attribute Evaluation

We verified the effectiveness of the learned attribute evaluator by the ranking accuracies. The ranking accuracy of each attribute is derived by 10-fold cross validation on the 150 labeled pairs in each primitive action, which is shown in Table 5. The result in the table demonstrates that our attribute evaluator, albeit learned only from relative information, has a high validity. In this experiment, only 3 primitive actions were considered here, i.e. L, T, and P. We combined segments of LM and UM with their corresponding subsequent operations of L, T and P, since LM and UM can be considered as the "preparation" step for the other operations. Some attributes are not considered intentionally for some actions (the "N/A" entries in Table 5) as it is not appropriate to assess the skills of the actions by these attributes. The result shows that the learned attribute learner achieves a significantly high accuracy for our defined skill attributes.

3.3. Instructive Video Evaluation

We compared our instructive video retrieval method with a baseline method that randomly selects one expert video clip of the primitive action. The comparison protocol is as follows. First, a query clip is selected from the database. Then, the recommended clips are obtained from both the proposed method and the baseline method. The two illustrative clips are paired in random order and presented to 8 human evaluators to judge which retrieved clip is more instructive. For each primitive operation, totally 60 queries are generated and the subjective evaluation result is summarized in Table 6. The "Instructive rate" shows the percentage of the retrieved videos are deemed

Table 6. Subjective evaluation result of the instructive video retrieval(%).

	Instructive rate	Comparative rate
L	93.3/83.3	50.0/40.0/10.0
T	96.7/73.3	63.3/26.7/10.0
P	95.0/76.7	60.0/26.7/13.3

instructive for the proposed/baseline approach. The "comparative rate" is the percentage of our proposed approach retrieving more/similar/less instructive videos than the baseline approach. The result shows that both methods present high instructive rate but the proposed method is persistently better than the baseline method. The result is especially satisfactory since the baseline method already employs expert videos, and thus our method is able to tell which expert video clip is more helpful to serve as an instructive reference. Since the proposed instructive video retrieval method is based on skill attribute analysis, this demonstrates the validity of the attribute learning.

4. CONCLUSION AND FUTURE WORK

In this paper, we presented a video-based surgical skill coaching system, aiming at providing weakness specific, skill superior and content similar instructive feedback. To build the system, we proposed a new problem defined as instructive video retrieval together with an effective framework to solve the problem by extending the idea of image attribute learning into video for skill understanding. In building the system, algorithmic innovations were made to incorporate domain knowledge and to handle practical difficulties arising from the real training platform. To our knowledge, this is the first video-based approach to delivering a systematic solution to the problem of automated skill coaching in simulation-based surgical training. Experiments with real world videos capturing the training sessions of resident surgeons have demonstrated the effectiveness of the idea and the key algorithms.

5. ACKNOWLEDGMENT

The work was supported in part by the National Science Foundation (NSF) and a grant from the U.S. Army Research Office (ARO). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of sponsors.

6. REFERENCES

[1] Dimitrios S. Alexiadis, Philip Kelly, Petros Daras, Noel E. O'Connor, Tamy Boubekeur, and Maher Ben Moussa, "Evaluating a dancer's performance using kinect-based skeleton tracking," in *Proc. of ACM MM'11*, 2011.

[2] S. Essid, D. Alexiadis, R. Tournemenne, M. Gowing, P. Kelly, D. Monaghan, P. Daras, A. Dremeau, and N.E. O'Connor, "An advanced virtual dance performance evaluator," in *Proc. of ICASSP'12*, 2012.

[3] Yu Jin, Xiaoxiang Hu, and GangShan Wu, "A tai chi training system based on fast skeleton matching algorithm," in *Proc. of ECCV'12*, 2012.

[4] K. Tervo, L. Palmroth, and H. Koivo, "Skill evaluation of human operators in partly automated mobile working machines," *Automation Science and Engineering, IEEE Transactions on*, Jan 2010.

[5] J. Rosen, B. Hannaford, C.G. Richards, and M.N. Sinanan, "Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills," *Biomedical Engineering, IEEE Transactions on*, May 2001.

[6] D. Coyle, *The Talent Code: Greatness Isn't Born. It's Grown. Here's How.*, Random House Publishing Group, 2009.

[7] Jiajun Liu, Zi Huang, Hongyun Cai, Heng Tao Shen, Chong Wah Ngo, and Wei Wang, "Near-duplicate video retrieval: Current research and future trends," *ACM Comput. Surv.*, Aug. 2013.

[8] Sara Memar, LillySuriani Affendey, Norwati Mustapha, ShyamalaC. Doraisamy, and Mohammadreza Ektefa, "An integrated semantic-based approach in concept based video retrieval," *Multimedia Tools and Applications*, 2013.

[9] Seung kook Jun, M.S. Narayanan, P. Agarwal, A. Eddib, P. Singhal, S. Garimella, and V. Krovi, "Robotic minimally invasive surgical skill assessment based on automated video-analysis motion studies," in *Proc. of BioRob'12*, June 2012.

[10] 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," *Biomedical Engineering, IEEE Transactions on*, March 2006.

[11] Antonio Criminisi, Jamie Shotton, and Ender Konukoglu, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," *Found. Trends. Comput. Graph. Vis.*, Feb. 2012.

[12] Ming-Kuei Hu, "Visual pattern recognition by moment invariants," *Information Theory, IRE Transactions on*, February 1962.

[13] Krishna Moorthy, Yaron Munz, Sudip K Sarker, and Ara Darzi, "Objective assessment of technical skills in surgery," *BMJ*, 10 2003.

[14] Richard Reznick, Glenn Regehr, Helen MacRae, Jenepher Martin, and Wendy McCulloch, "Testing technical skill via an innovative bench station examination," *The American Journal of Surgery*, 1997.

[15] Jeffrey D. Doyle, Eric M. Webber, and Ravi S. Sidhu, "A universal global rating scale for the evaluation of technical skills in the operating room," *The American Journal of Surgery*, 2007.

[16] D. Parikh and K. Grauman, "Relative attributes," in *Proc. of ICCV'11*, Nov 2011.