Instructive Video Retrieval Based on Hybrid Ranking and Attribute Learning

A Case Study on Surgical Skill Training

Lin Chen, Peng Zhang, Baoxin Li
Computer Science and Engineering
Arizona State University, Tempe, AZ, 85281
{lchen109, pzhang41, baoxin.li}@asu.edu

ABSTRACT

Video-based systems have been increasingly used in various training tasks in applications like sports, dancing, and surgery. One key task to add automation to such systems is to automatically select reference videos for a given training video of a trainee. In this paper, we formulate a new problem of instructive video retrieval and propose a solution using both attribute learning and learning to rank. The method first evaluates a user's skill attributes by relative attribute learning. Then, the most critical skill attribute in need of improvement is selected and reported to the user. Finally, a hybrid ranking learning to rank method is employed to retrieve instructive videos from a dataset, which serve as reference for the user. Two main technical problems are solved in this method. First, we combine both skill and visual feature to characterize skill superiority and context similarity. Second, we propose a hybrid ranking approach that works with both pair-wise and point-wise labels of the data. The benefit of the proposed method over other heuristic methods is demonstrated by both objective and subjective experiments, using surgical training videos as a case study.

Categories and Subject Descriptors

I.4 [**Image Processing and Computer Vision**]: Application

Keywords

Video Retrieval; Attribute Learning; Learning to Rank

1. INTRODUCTION

Video-based training systems have found many applications in sports, dancing, and surgery, etc., where cameras are used to record and/or monitor performance of the trainees. Wherever expert supervision is scarce/expensive, a system that can provide automatic feedback based on a trainee's performance would be desired for a trainee's self-improvement. One important type of feedback is to provide verbal instructions based on analysis of the trainee's performance *and* illustrative videos of other higher-skilled performers. We term this as instructive video retrieval.

MM'14, November 03–07, 2014, Orlando, FL, USA.

http://dx.doi.org/10.1145/2647868.2655050.

Instructive video retrieval is a new problem that is different from traditional video retrieval tasks such as nearly-duplicated video retrieval and concept retrieval.In this new problem, we need to focus on whether the retrieved video is helpful for skill improvement. To our knowledge, this is a new research effort with little prior art. One challenge of such task lies in the inherent vagueness and subjectivity of the term "instructive", which makes it difficult to formulate a retrieval cost function. In this work, the instructiveness of a video is defined by considering three criteria: specificity, superiority and similarity. Specificity requires that the illustrative video should be selected w.r.t. a trainee's skill weakness and a corresponding verbal instruction would be provided telling why this video is recommended. Superiority refers to that the illustrative video should represent a better skill on the weakness aspect (attribute) of the trainee. Similarity means that the illustrative video should have a similar operation context (e.g., performing similar actions) to the trainee's video so that it is easy for the trainee to figure out how to imitate and improve skills.

Considering these criteria, in this paper, we design an instructive video retrieval method based on attribute learning and learning to rank, using surgical training as a case study. Our hybrid ranking approach combines both pair-wise (relative) and point-wise (binary) data, making it flexible in handling variable levels of availability of labeled training data in a practical application. The major contribution of this paper is three-fold. First, we formulate the new problem of instructive video retrieval and propose a learning-based solution. Second, in the chosen case study, both skill attributes and operation context features are designed to implement the general approach for delivering good performance. Last, we present a hybrid ranking SVM to take advantage of both pair-wise and point-wise labels for more effective ranking under sparse and noisy labels.

2. RELATED WORK

Human skill assessment is a research topic with a long history originated from psychology where skill is defined as the relation between task difficulties and the resources paid for and quality gained from the task[3]. There are generally two ways for automatic skill evaluation. One is to build a model for different skill levels, and use the model distance from the user's performance for skill assessment, e.g., [2]. This method provides an overall skill assessment but it is not

specific enough to tell trainees how to improve. The other is to provide performance metrics based on motion feature statistics, e.g., [1]. But these mere metric are difficult to use for coaching. Our attribute learning method adopted skill attributes from domain knowledge[2]. Therefore, it can not only evaluate different skill aspects, but also generate instructions understandable to trainees. Moreover, we can retrieve an illustrative video via attribute analysis, which is one big step further than only skill evaluation.

Attribute learning maps low-level image feature to intermediate visual attributes, e.g. fur, color, and stripe, rather than high-level category labels, e.g. zebra. The introduction of semantic attributes greatly improves human-computer-interaction capability and boosts many learning applications with human involved [6]. Our work is related to multi-attribute image retrieval, since we will consider all skill attributes for instructive video selection. The multi-attribute fusion function can either be designed from heuristics, e.g., L1 norm of matching score on each attribute[8], or be learned from model assumption, and the model can be either generative graphical model, e.g., Bayesian network[9], or discriminative function, e.g., SVM[10]. We formulate our problem in a discriminative function since generative models requires more training samples while heuristic fusion is difficult due to the vagueness of instructiveness.

Learning to rank [4] is different from traditional heuristic method where ranking metrics is learned from labels. There are generally three types of labels for LR problem: point-wise, pair-wise and list-wise. However, there are new challenges in the instruction video retrieval problem. First, due to the high cost of labeling real training videos by expert surgeons, the training labels are typically very sparse. Second, the labels may be noisy because of the vagueness in instructiveness concept and the subjective nature of the problem. To this end, we present a hybrid-ranking SVM method to take advantage of both point-wise and pair-wise labels. There are other hybrid methods. For example, [5] uses isotonic regression to optimize pair-wise margin among samples of different relevance levels where point-wise relevance constraint can be further incorporated. In [7], pair-wise margin between different relevance levels are formulated in list-wise NDCG cost function and the function can be solved by unconstraint optimization. These methods only use point-wise label and their hybrid only appears in cost function. In contrast, our method makes use of both point-wise and pair-wise labels and we can thus even differentiate pairs on the same relevance level.

3. SURGICAL TRAINING TASK

The Fundamentals of Laparoscopic Surgery (FLS) training box has been widely used in minimally-invasive surgery training. The system is a box simulating the human body and a trainee uses tools going into the box through two small holes to perform surgical actions. The trainees (Fig. 1 Left) can only watch the inside of the box through a monitor, which is captured by an on-board camera (Fig. 1 Right).

In this training procedure, a trainee is required to lift one of the six objects with a grasper in one hand, transfer the object to another hand, and then place it 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. The peg transfer operation consists of 3 primitive actions in Tab. 1. Ideally, we should perform each primitive action once to finish one peg transfer cycle. For six objects being transferred from left to right and backward, there are totally 12 cycles in one training session.

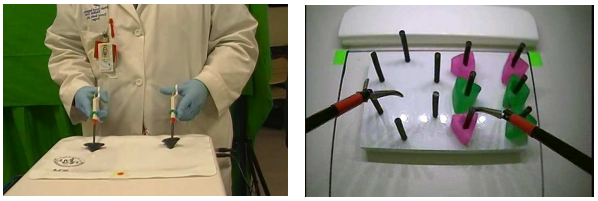

Figure 1: Illustrating the FLS system.

Table 1: Primitive Actions of Peg Transfer

Name	Description
Lift	Move grasper to a peg with object and raise the object off
Transfer	Move the two graspers together and pass the object from one grasper to another
Place	Move grasper to an empty peg and release the object

Surgical skill can be evaluated by a set of attributes. Following the popular Global Rating Index for Technical Skills[2], we adopt several important skill attributes most relevant to the peg transfer task for skill evaluation and instructive video retrieval. These attributes are listed in Tab. 2.

4. METHODOLOGY

Given a clip of trainee's operation, our task is retrieving and recommending an instructive video clip for skill improvement. We propose to solve this task in three steps as elaborated below.

4.1 Critical Skill Attributes Selection

To retrieve instructive videos for a trainee's skill improvement, we must first figure out the operation weakness. A lowest attribute value does not always mean the most urgent attribute in need of improvement. The weakest attribute should be the one that the trainee did poorly while most other people are significantly better.

We first perform normalization across different attributes so that the values of the skill attributes would reflect their respective urgency to be improved, shown in Eq (1):

$$a'_{k,i} = P_k(a \geq a_{k,i} + S_k), (1 \leq k \leq K) \quad (1)$$

where $a_{k,i}$ is the attribute value of the k-th skill attribute $A_k(1 \leq k \leq K)$ for the i-th video V_i. $P_k(\cdot)$ is the probability distribution of attribute A_k on a pre-defined video database and the S_k is the significance threshold of A_k. In fact, Eq (1) is an "inverse" normalization into attribute importance, because higher skill attribute value represents less urgent attribute to improve and will leads to lower normalization value. The success of (1) lies in a well-constructed database with video clips from people of different skill levels.

Table 2: Skill Attributes for Surgical Skill

Attribute	Define
Time and motion	Unnecessary moves and time cost
Flow of operation	Moves smoothly without stop
Bimanual dexterity	Cooperation between hands
Instrument handling	Tentative operations and errors

The second processing step is attribute comparison. Given a user's performance video V_i, and a retrieved video V_j, we measure how much V_j can help V_i on the attribute A_k. This is called the utility of V_j to V_i on A_k, and is measured by the difference between the normalized attribute values in (1) as

$$u_{k,i,j} = (a'_{k,j} - a'_{k,i})/(a'_{k,i}), (1 \leq k \leq K) \quad (2)$$

4.2 Skill and Operation Context Features

Let $\boldsymbol{f}_{q,d}$ be the ranking feature derived from the pair of a trainee's video V_q and a reference video V_d. $f_{q,d}$ should consist of both skill context and operation context to embody the specificity, superiority, and similarity.

Specificity is determined by the normalized attribute importance in (1). Superiority principle can be described by the difference between the normalized attribute value of V_q and V_i in (2). So we define the skill context feature as:

$$\boldsymbol{f}^{(s)}_{q,d} = [\boldsymbol{a}_q, \boldsymbol{a}_d, \boldsymbol{u}_{q,d}] \quad (3)$$

where $\boldsymbol{a}_q/\boldsymbol{a}_d$ are the normalized attributes value vector of V_q/V_d as $\boldsymbol{a}_q = [a'_{k,q}]$ and $\boldsymbol{u}_{q,d}$ is the utility vector of V_d to V_q as $\boldsymbol{u}_{q,d} = [u_{k,q,d}]$, where $k = 1, \cdots, K$.

Similarity principle means V_d should have a similar operation context as V_q, e.g., using the same hand and with similar surroundings. To measure this, first we divide the operation context of an image I into blocks $B_n, n = 1, \ldots, N$. Each B_n is described by a block context vector $\boldsymbol{b}_n = [c_1, \ldots, c_M]$ where c_m is the area portion of the m-th region category, e.g., object and peg, in block B_n. Since FLS box is a controlled environment with significant color difference among region categories, the area portion c_m can be obtained by color segmentation and tracking. Second, we describe the similarity between a frame I in V_q and its peer frame I' in V_d based on block similarity. Each block may have different importance for the similarity by its relative position to the grasper. To address this, we take the block containing the grasper tip of I as the reference center, and order all blocks by the distance and angle to the center. Suppose $B_n, n = 1, \ldots, N$ are already ordered by the above rule, so the first block B_1 contains the reference center. The blocks B'_n of frame I' will follow the same order so that B'_n and B_n are of the same location. The block context vector $\boldsymbol{b_n}$ already conveys block context, since it is derived from object segmentation. So the block similarity can be directly defined as the inner product of block context vectors, and the similarity between I and I' is defined as:

$$s(I; I') = [\langle b_1, b'_1 \rangle, \ldots, \langle b_N, b'_N \rangle]. \quad (4)$$

Then the operation context feature of video pair (V_q, V_d) can be represented as

$$\boldsymbol{f}^{(v)}_{q,i} = [\boldsymbol{s}(I_0; I'_0), \boldsymbol{s}(I_1; I'_1)], \quad (5)$$

where $I_0/I_1(I'_0/I'_1)$ is the start/end frame of $V_q(V_d)$.

The final ranking feature of video pair (V_q, V_d) is:

$$f_{q,d} = [f^{(s)}_{q,d}, f^{(v)}_{q,i}] \quad (6)$$

which will guarantee specificity, superiority, and similarity principles for the following learning to rank.

4.3 Hybrid Ranking SVM

With the selected attribute and extracted ranking features, we now propose to employ hybrid ranking SVM to take advantage of both point-wise and pair-wise labels. Pair-wise ranking SVM, also called ranking SVM[6], tries to find a projection vector w to satisfy the pair label under maximal margin assumption. Point-wise ranking SVM tries to find a projection vector and a set of intersections related to different relevance level in maximal margin spirit, e.g. CO SVM[4]. The pair-wise and point-wise methods have their advantages and disadvantages. The pair-wise label is very expensive, but it can provide precise order between two samples. The point-wise label is less expensive, but it can't rank samples of the same relevance level. Since video labeling is very expensive, the labeling can be very sparse. It is better to combine the two label types for more effective ranking. For example, we can include the pair-wise label between relevant samples to augment the point-wise label.

The direct combination the cost function and constraints in [6] and [4] is not desirable, because it fix the margin thresh of point-wise and pair-wise sample to equal value, i.e. 1. Our idea is to introduce an extra variable K as the margin thresh, and reformulate SVM as below:

$$\begin{aligned} \min_{\boldsymbol{w},\varepsilon,w_0,\gamma,K} & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C_1 \sum_{i=1}^{m} \varepsilon_i + C_2 \sum_{i=1}^{m} \gamma_i \\ s.t. \quad & y_i \cdot \langle w, \boldsymbol{x}_i^{(1)} - \boldsymbol{x}_i^{(2)} \rangle \geq 1 - \varepsilon_i \\ & z_i \cdot (\langle \boldsymbol{w}, x_i \rangle + w_0) \geq K - \gamma_i \\ & \gamma_i \geq 0, \varepsilon_i \geq 0. \end{aligned} \quad (7)$$

where $\boldsymbol{w}$ is the ranking model; $y_i = 1(-1)$ means $\boldsymbol{x}_i^{(1)}$ is superior (inferior) to $\boldsymbol{x}_i^{(2)}$; $z_i = 1(-1)$ means x_i is instructive/non-instructive; ε and γ are slack variables. This problem can be solved by quadratic programming.

Table 3: The ranking result of pair-wised, point-wised and hybrid ranking SVM in simulated data.

Noise	Cosine			Kendall		
	Pair	Point	Hybrid	Pair	Point	Hybrid
1%	0.39	0.47	0.49	0.25	0.32	0.33
10%	0.28	0.40	0.42	0.18	0.26	0.28
30%	0.15	0.19	0.24	0.09	0.12	0.15

Table 4: The classification and ranking accuracy (%) on surgical video data of three primitive actions.

	Lift	Transfer	Drop
Point/Hybrid	78.1/80.7	83.7/84.4	83.4/88.4
Pair/Hybrid	75.7/76.3	90.0/89.8	89.3/90.4

5. EXPERIMENTS

We first show the efficiency of hybrid-ranking SVM by both simulation and real surgical video data. Then we show the effectiveness of our instructive video retrieval system. For each of the three primitive actions, our surgical video dataset consist of 240 video clips on different skill levels, i.e. from novice to expert. So there are totally 720 video clips.

5.1 Simulation Experiment

We first compared our Hybrid-based SVM approach (7) with two baseline approaches of pair-wise method[6] and point-wise method[4]. The simulated dataset are uniformly generated with 150 point-wise samples and 150 pairs-wise samples, whose labels are decided by a randomly generated projection vector w. Additionally, noise is added into the la-

bel for robustness test. Tab. 3 shows the experiment result of the three approaches when noise are added by 1%, 10% and 30%. Two measures are adopted to compare the performance. The first measure is the direction cosine similarity between the projection vector of trained model and the originally generated one; the other measure is the Kendall rank correlation coefficient of the projection result. Result in Tab. 3 shows that the point-wise label is much more efficient than pair-wised label in learning the true projection angle. However, the hybrid-ranking SVM can still improve the ranking accuracy by combining both point-wise and pair-wise labels. And the improvement is more significant with the increase of noise level.

5.2 Surgical Data Classification and Ranking

The training data were labeled in the following procedure. Given a random action clip as query, we randomly pick another two clips from the clips with significant higher attribute values. Then the expert surgeon would evaluate if the returned clips is instructive or not instructive. If both of the two clips are instructive, the expert would further provide a pair-wise label indicating which clip is more instructive. The labeling process follows the three criterions discussed above. For each of the 3 primitive actions, we generated 30 queries and 8 pairs of clips for each query.

Similarly, we illustrate the benefit of hybrid ranking SVM by comparison with both point-wise and pair-wise SVM. We train hybrid ranking SVM with both the point and pair labels acquired in the above process, while the point-wise and pair-wise SVM are trained with their corresponding labels. Tab. 4 shows the classification and the ranking accuracies of the hybrid-ranking SVM based on the combination of both point and pairwise labels, compared with the purely point-wise and pair-wise based approaches. Each entry shows the accuracy in lift/transfer/drop action. The results show that our hybrid approach provides better accuracy than the two baseline methods.

5.3 Instructive Video Selection

Finally we compare our video with a baseline method that randomly selects one expert video clip of the primitive action. The subjective comparison protocol is as follows. We recruited 6 subjects who had no prior knowledge on the dataset. For each subject, 10 clips for each primitive action are given as query. For each query, both our hybrid ranking SVM and the baseline method will return an instructive clip. So there are totally 60 queries for each of the 3 primitive actions. For each returned video pairs, the subjects need to evaluate whether the video is instructive in improving the operation in the query video, and which one is more instructive. The comparative result in Tab. 5 shows that our hybrid ranking SVM performs much better than the baseline method in selecting instructive videos. The first row shows the percentage of instructive coaching videos returned by our hybrid/baseline approaches. The second row shows the percentage that the returned video of our approach is more/equal/less instructive than the baseline approach. Result shows that, the returned videos by our approach are almost all instructive, apparently higher than the baseline approach. Even in the case that both returned videos from two approaches are instructive, our approach would recommend more instructive videos than the baseline approach by large majority.

Table 5: Subjective evaluation results (%) of instructive video selection.

	Lift	Transfer	Drop
Instructive	98.1/74.8	98.3/83.1	99.2/73.4
Superiority	61.1/21.3/17.6	73.2/15.5/11.3	68.1/16.1/15.8

6. CONCLUSIONS

We formulated a new problem of instructive video retrieval and developed a solution in the case study of surgical training video. In defining the problem based on three criteria, we allow the otherwise abstract and subjective task to be attacked by relative attribute learning. To facilitate a realistic solution, considering the poor availability of labeled data, we proposed a new hybrid ranking method to approximate instructive score from both attribute skill and visual operation context features which are designed to embody the three criterions, taking advantage of both point-wise and pair-wise labels. Stimulation and real data experiments demonstrated the approach is effective and thus can be a promising solution to the instructive video retrieval problem.

7. ACKNOWLEDGMENTS

The work was supported in part by National Science Foundation (NSF) (Grant 0904778). The views expressed in this material are those of the author(s) and do not necessarily reflect the views of NSF.

8. REFERENCES

[1] D. S. Alexiadis, P. Kelly, P. Daras, N. E. O'Connor, T. Boubekeur, and M. B. Moussa. Evaluating a dancer's performance using kinect-based skeleton tracking. In *ACM MM*, 2011.

[2] J. D. Doyle, E. M. Webber, and R. S. Sidhu. A universal global rating scale for the evaluation of technical skills in the operating room. *The American Journal of Surgery*, 2007.

[3] P. M. Fitts and J. R. Peterson. Information Capacity of Discrete Motor Responses. *J Exp Psychol*, Feb. 1964.

[4] T.-Y. Liu. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, Mar. 2009.

[5] T. Moon, A. Smola, Y. Chang, and Z. Zheng. Intervalrank: Isotonic regression with listwise and pairwise constraints. In *WSDM*, 2010.

[6] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, Nov 2011.

[7] C. Renjifo and C. Carmen. The discounted cumulative margin penalty: Rank-learning with a list-wise loss and pair-wise margins. In *MLSP*, Sept 2012.

[8] W. Scheirer, N. Kumar, P. N. Belhumeur, and T. E. Boult. Multi-attribute spaces: Calibration for attribute fusion and similarity search. In *CVPR*, June 2012.

[9] W. Scheirer, N. Kumar, K. Ricanek, P. Belhumeur, and T. Boult. Fusing with context: A bayesian approach to combining descriptive attributes. In *IJCB*, Oct 2011.

[10] B. Siddiquie, R. Feris, and L. Davis. Image ranking and retrieval based on multi-attribute queries. In *CVPR*, 2011.