Human Action Recognition in Table-top Scenarios

An HMM-based Analysis to Optimize the Performace

Ramana, Pradeep Kumar Reddy; Grest, Daniel; Krüger, Volker

Published in:
Computer Analysis of Images and Patterns

Publication date:
2007

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Ramana, P. K. R., Grest, D., & Krüger, V. (2007). Human Action Recognition in Table-top Scenarios: An HMM-based Analysis to Optimize the Performace. In W. G. Kropatsch, M. Kampel, & A. Hanbury (Eds.), *Computer Analysis of Images and Patterns: 12th International Conference, CAIP 2007, Vienna, Austria, August 27-29, 2007, Proceedings* (pp. 101-108). Springer. Lecture Notes in Computer Science Vol. 4673

Human Action Recognition in Table-top Scenarios : An HMM-based Analysis to Optimize the Performace

Pradeep K Reddy, Daniel Grest, Volker Krueger

Computer Vision and Machine Intelligence,
Copenhagen Institute of Technology,
Aalborg University, Denmark.
{pradeep,dag,vok}@cvmi.aau.dk

Abstract. Hidden Markov models have been extensively and successfully used for the recognition of human actions. Though there exist well-established algorithms to optimize the transition and output probabilities, the type of features to use and specifically the number of states and Gaussian have to be chosen manually. Here we present a quantitative study on selecting the optimal feature set for recognition of simple object manipulation actions pointing, rotating and grasping in a table-top scenario. This study has resulted in recognition rate higher than 90%. Also three different parameters, namely the number of states and Gaussian for HMM and the number of training iterations, are considered for optimization of the recognition rate with 5 different feature sets on our motion capture data set from 10 persons.

Keywords Hidden Markov model, Action Recognition, Optimization

1 Introduction

Extensive research has been carried out in the past to understand and recognize human actions. Typical applications of the human activity recognition are surveillance, human-robot interaction, imitation learning[1, 3]. In surveillance, it is important to detect abnormal and suspicious actions. In robotics community, huge body of research is in place to imitate human through demonstration, imitation and learning. Research on imitation learning shows that actions can be clearly segmented into atomic action units [8, 9]. This gives a way to recognize human actions through interpretation of its atomic units [10, 4], such as *pointing, grasping, and rotating* in case of table-top scenarios. Though the scenario is simple, due to the similarity of different actions, the recognition task is difficult. Due to the simplicity of the scenario, the action recognition study on this scenario lets us gain understanding of the human motion and gives insights as to how to tackle the problem of action recognition in complex scenarios.

Considerable research in computer vision community has focused on cyclic motions, such as walking or running [1] and Jenkins et. al [10] came up with a system to extract the behavior vocabularies for the problem of complex task learning. [11] presents a learning system for one and two hand motions to compute the hand motion trajectory that optimizes the imitation performance given the constraints of the body of the robot.

Here we present a quantitative study to optimize the recognition performance of the simple human actions, like *pointing, grasping, and rotating* performed in a table-top scenario. Our action dataset consists of 10 persons, in contrast to many other studies with 1/2 persons [1, 6, 7]. We use the Hidden Markov Model(HMM) to learn the motion model. It defines the joint probability distributions over observations and the states of the model. An HMM does time-warping to some extent implicitly in the cases of motion sequences, which differ slightly in their execution speed. Hence it is well-suited to our action dataset consisting of 10 persons with differing execution paces. Our study is based on about 890 sequences, each sequence with an average of 115 poses, totaling 102350 poses.

Though there exist well-established algorithms, such as the Baum-Welch algorithm, to optimize the transition and output probabilities of a HMM, the structure of the HMM and the its inputs have to be chosen manually. The feature set used to train the HMM is really important, as an inappropriate feature set could lead to low performance of the system and making the system not being able to capture the discriminative features among different actions. Apart from the feature set, the number of states for the HMM and the number of Gaussian for each state has to be given by hand. Moreover the number of training iterations for the HMM also need to be set beforehand. These parameters are usually empirically determined, but this task is often laborious and time consuming in case of big datasets. Hence this study saves a lot of effort and time for subsequent researchers, working in similar scenarios. Moreover this scenario is gaining importance in imitation learning.

Perhaps the work most similar to our work presented here is that of Guenter et. al [12]. They have presented simple algorithms to optimize the number of states, training iterations and Gaussians for the HMM in the context of handwritten word recognition task. Their work merely finds parameters with optimal performance, without any studies on different feature sets. Here we perform an extensive evaluation of different feature sets and present results on our action dataset, pointing out the feature set with best classification rate. This evaluation is necessary to build a system with the best performance. Enroute to finding the optimal feature set, the optimal values for different parameters of HMM will also be determined. Recent research [5] shows that one can do motion tracking even from a single view. So this study could be combined with motion capture to recognize the actions in real-time.

The paper is organized as follows: Section 2 briefly describes the action recognition through Hidden Markov models and Section 3 describes the motion capture setup and collection of training data. In Section 4, we describe the different feature sets used for the evaluation. Section 5 describes the quantitative evaluation of different feature sets and we conclude in Section 6.

2 HMM based Human Action Recognition

L. Rabiner [2] gives an excellent tutorial on how to use HMMs. Given an observation sequence $O = O_1 O_2 \ldots O_T$ and a model $\lambda = (A, B, \Pi)$, the probability of the observation sequence O, given the model $\lambda, P(O \mid \lambda)$ is calculated by enumerating over all possible state sequences of length T, $Q = q_1 q_2 \ldots q_T$. For an observation sequence O and state sequence Q,

$$P(O \mid Q, \lambda) = \prod_{t=1}^{T} P(O_t \mid Q_t, \lambda) \tag{1}$$

and then the probability of the observation sequence given the model, irrespective of the state sequence, is obtained by summing the probability over all the possible state sequences:

$$P(O \mid \lambda) = \sum_{allQ} P(O \mid Q, \lambda) P(Q \mid \lambda) \tag{2}$$

Then we follow the forward-backward algorithm [2], to compute the likelihood of the observation sequence given the HMM. we train one HMM for each class of sequences. To recognize an observation sequence from a dataset, we calculate the likelihood of this sequence from all the trained HMMs. The HMM that gives maximum likelihood represents the observation sequence. If this HMM and the observation sequence belong to the same action class, then we say we could correctly recognize the sequence.

3 Our Action Dataset

Our dataset consists of four actions: *Pointing (P), Grasping (G), Rotating (R) and Displacing(D)*, performed by 10 persons. In this work we consider three different actions: *P, G and R.* Each action is performed

1. in three different directions (to the right, to the front and to the left)
2. with two different heights and
3. with two different distances.

Each person repeated each action 5 times for all combinations of the above settings. With 10 persons, this gives rise to good amount of training examples with respect to *variability in execution speed, direction, height, scale and arm lengths.* This is essential in order to capture the key features of a particular action invariant to aforementioned properties.

Each person has 7 electromagnetic sensors. They are located at the upper back of the torso, shoulder, elbow, wrist, hand, thumb and the index finger, as depicted in Figure 1. The measurements were acquired with Motionstar of Ascesion [14]. For each sequence, the 3D coordinates of the 7 sensors were recorded at 25 fps, starting from a nearly vertical position (rest pose, Figure 2). The sequence ends also in the rest pose. Figure 2 shows a person performing the action

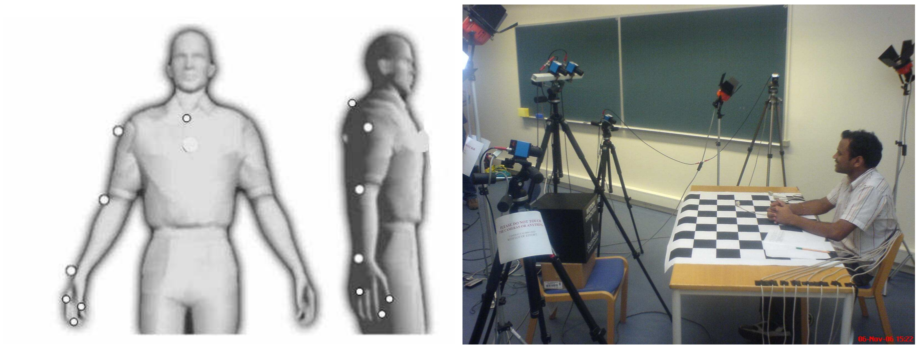

Fig. 1. Locations of the electromagnetic sensors on the body of the person and the Setup used to record both markers and video data

Rotating. The scenario was also video-recored by 4 well-calibrated cameras simultaneously with the marker data. The video data could be used for motion tracking etc. The whole setup used is shown in Figure 1.

4 Feature Sets

Each feature set is a sequence of features for all poses. For all the sequences, the 3D position of the torso is considered to be the origin for each pose. It is subtracted from the positions of all the sensors, so that the data used for recognition is independent of the position of the person performing the action. In order to determine the best feature set, we've extracted the following feature sets appropriate for the one-arm movements in a table-top scenario:

1. Joint Angles **(JA)**:This consists of the angles at the shoulder and at the elbow between the lower arm and the upper arm. Knowing the problem of singularity i.e. 0 and π are same, we use the up and direction vectors in the shoulder coordinate frame, obtained from the kinematic chain built for the right arm for each pose. This is invariant to the arm length and can handle directions easily.
2. Direction vector from Torso to Wrist (**T2W**)
3. Direction vector from Shoulder to Wrist (**S2W**)
4. **T2W** + *Grasp Distance* **GD**: Here we define Grasp Distance to be the distance between the sensors on the thumb and the index finger. It is intuitive to see that, GD is discriminative feature between Grasping and the rest. This is confirmed by the experiments.
5. **S2W + GD** Direction vector from Shoulder to Wrist and the grasp distance.

5 Quantitative Analysis

In order to learn and recognize different actions, we trained one HMM for each class of sequences. Each time three quarters of sequences from each class are

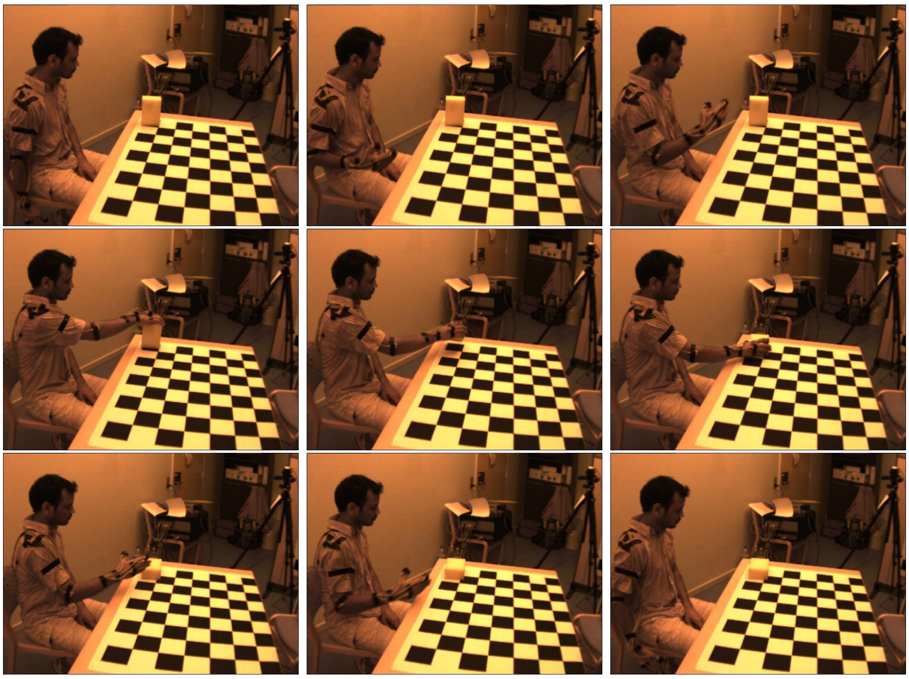

Fig. 2. Sequence of Images of a Person Performing the action *Rotating*

reserved for training and the fourth quarter is used to test the performance of trained HMMs. The recognition is done using Maximum Likelihood approach as follows. The likelihood of test sequence against all the trained HMMs are calculated and the HMM with maximum likelihood represents the test sequence.

We use the Hidden Markov Model Toolbox for Matlab, developed by Kevin Murphy [15] for training and testing. Having chosen different feature sets, its important to set the suitable architecture for the HMM for each of these feature sets to maximise the classification rate. This means that the optimal values for different parameters of the HMM have to be determined. From the preliminary trials on the dataset, we learned that the following ranges of parameters are producing recognition rates(RR) within acceptable range:

- Number of States (**Q**): from 5 to 40 in steps of 5
- Number of Gaussian (**G**) for each state: 10, 20 and 30
- Number of Training Iterations (**I**): 10, 15 and 20

With these ranges for different parameters, we use a simple *pick-the-best* algorithm to find the optimal values for the parameters corresponding to each feature set. For each feature set, we enumerated over the values of each parameter in the above ranges and the performance is noted. The set of parameters corresponding to the maximum recognition rate is noted for each feature set. Though the algorithm is straightforward, it is important to make an extensive evaluation over the above ranges, to make sure that we do not miss the best combination. Each run of the algorithm, with one set of parameters takes about 25 minutes on an average with a dual core 2GHz computer.

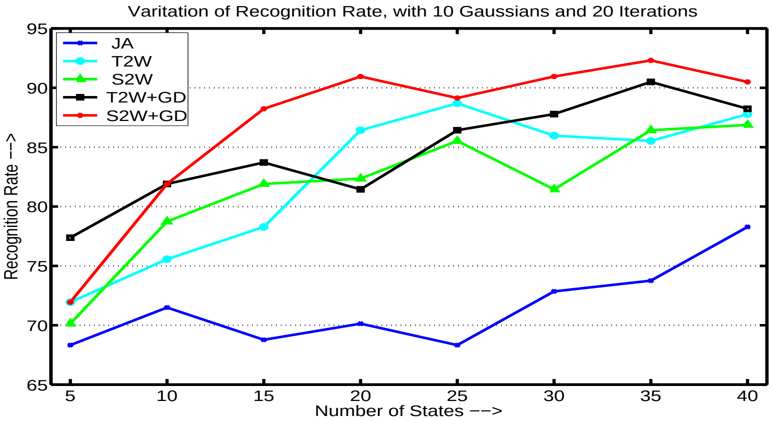

Fig. 3. Variation of RR over number of states, with 10 Gaussian and 20 Iterations

Table 1. Best Recognition Rates for Different Feature Sets

Feature Set	Best Recognition Rate	Optimal Q, G, I
JA	80.09 %	40, 20, 20
T2W	88.68 %	25, 10, 20
S2W	87.78 %	30, 10, 15
T2W+GD	90.49 %	35, 10, 20
S2W+GD	94.12 %	35, 10, 15

The best recognition rates and the optimal values for the HMM parameters are given in Table 1. From Table 1, it is clear that GD definitely improved the performance of the system. **S2W+GD** has optimal performance. One reason is be that the motion of shoulder joint for different persons is different for the same action. So it helped the HMM to learn the key features common to the action and to dispose the variant features, and obtain better discrimination among different actions. This study also shows that inclusion of elbow motion(in feature set **JA**) degraded the classification rate, rather than improving it. It was also shown by Vicente et. al. [13]. From Table 1, we can also see that the combination of 10 Gs and 15 Is, 10 Gs and 20 Is have the best performance. Figure **??** show the variation of the recognition rate over the number of states with these combinations. We can see that the performance increases with the number of states gradually for many of the feature sets.

Apart from the best performance, the average performance of the particular feature set is important. This is because it can be expected, that a feature set with higher average performance and low sensitivity to HMM parameters would perform equally well on another datasets in a similar scenario. The mean, maxi-

Table 2. Statistics of the Recognition Rates for All Feature sets

Feature Set	Mean RR	Max. RR	Min. RR	Std.Dev
JA	71.23 %	80.09 %	57.92%	5.02
T2W	79.27 %	88.68 %	56.56%	6.37
S2W	76.99 %	87.78 %	33.48%	9.37
T2W+GD	82.13 %	90.49 %	58.37%	6.13
S2W+GD	84.16 %	94.12 %	45.70%	9.54

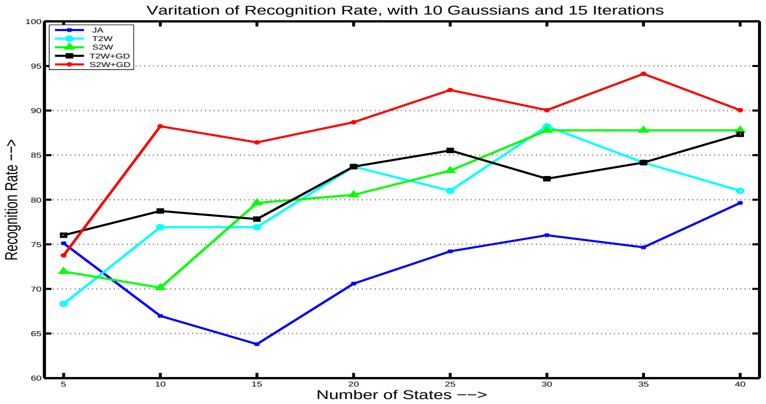

Fig. 4. Variation of RR over number of states, with 10 Gaussian and 15 Iterations

mum, minimum and the standard deviation of recognition rate over all the trials for the five feature sets are given in Table 2. The the standard deviation of the feature set **S2W+GD** is higher than than that of the feature set **T2W+GD**. But **S2W+GD** dominates **T2W+GD** clearly in mean and best recognition rates. So we can say that the feature set **S2W+GD** is very optimal and is suitable for movements in table-top scenarios. So the optimal set of parameters are determined to be 35 states for the HMM and 10 Gaussians for each state and 15 iterations for the training.

6 Conclusions

A quantitative study on optimizing the performance of the human action recognition system in table-top scenarios is presented. A set of simple object manipulative actions are used to find the best feature set and optimal values for the HMM parameters. These set of actions could be classified most accurately with the direction vector from shoulder to wrist and the grasp distance. The high average performance of this feature set shows that this is less sensitive to HMM parameters. So it is expected to perform equally well on different datasets in a similar scenario. Our study also shows that inclusion of elbow motion degraded the classification rate, rather than improving it. We are building an online task recognition system, integrating this with a motion capture system.

Acknowledgment This work has been supported by PACO-PLUS, FP6-2004-IST-4-27657.

References

1. J.K. Aggarwal, Q. Cai, Human motion analysis: A review, Computer vision and Image Understanding:CVIU, vol 73, no.3, pp 428-440,1999.
2. L. Rabiner, A tutorial on Hidden Markov models and selected applications in speech recognition, Proc. of the IEEE, 77(2):257-285,1989.
3. Moeslund, T. B., Hilton, A., and Krueger, V. 2006. A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding. 104, 2 (Nov. 2006), 90-126.
4. Volker Krueger, Daniel Grest. Using Hidden Markov Models for Recognizing Action Primitives in Complex Actions, Scandinavian Conference on Image Analysis (SCIA). 2007.
5. Daniel Grest, Reinhard Koch, Volker Krueger, Single View Motion Tracking by Depth and Silhoutte Information, Scandinavian Conference on Image Analysis (SCIA). 2007.
6. Campbell, L. and A. Bobick. Recognition of human body motion using phase space constraints. In International Conference in Computer Vision, pp. 624-630 Cambridge MA, 1995.
7. A. Billard, S. Calinon, F. Guenter, Discriminative and adaptative imitation in uni-manual and bi-manual tasks, in Robotics and Autonomous Systems, Vol. 54, pp. 370-384, 2006.
8. Mataric, M. J. Sensory-motor primitives as a basis for imitation: linking perception to action and biology to robotics. In Imitation in Animals and Artifacts, K. Dautenhahn and C. L. Nehaniv, Eds. MIT Press, Cambridge, MA, 391-422, 2002.
9. D. Newtson, Gretchen Engquist, Joyce Bois, The objective basis of behavior unit, Journal of Personality and social psychology, vol 35, no. 12, pp. 847-862, 1977.
10. O.C. Jenkins and M.J. Mataric, Performance derived vocabularies: Data-driven acquisition of skills from motion, International Journal of Humanoid Robotics, vol. 1, pp. 237-288, June 2004.
11. S. Calinon, A. Billard, and F. Guenter,Discriminative and adaptative imitation in uni-manual and bi-manual tasks, in Robotics and Autonomous Systems, vol. 54, 2005
12. S. Guenter and H. Bunke, Optimizing the Number of States and Training Iterations and Gaussians in an HMM-based Handwritten Word Recognizer, Proceedings. Seventh International Conference on Document Analysis and Recognition, Pages: 472-476, vol.1, Aug. 2003.
13. Isabel Serrano Vicente, Danica Kragic, Learning and Recognition of Object Manipulation Actions Using Linear and Nonlinear Dimensionality Reduction, 15th IEEE Int. Symp. on Robot and Human Interactive Communication(RO-MAN) , 2007. Submitted.
14. Ascension. Motion Star Real-Time Motion Capture. `http://www.ascension-tech.com/products/motionstar_10_04.pdf`.
15. Kevin Murphy, Hidden Markov Model (HMM) Toolbox for Matlab, 1998. `http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html`