Trajectory classification using HMMs

Jozef Mlích, Pavel Zemčík, Leoš Jiřík
imlich@fit.vutbr.cz ,zemcik@fit.vutbr.cz, xjirik02@stud.fit.vutbr.cz
Department of Computer Graphics and Multimedia
Faculty of Information Technology, Brno University of Technology
Božetěchova 2, Czech Republic, 612 66, Brno

ABSTRACT

This paper focuses on evaluation of motion of objects through classification of their trajectories. The objects used for evaluation of the presented method are objects tracked in video sequence but the method is quite general and can be used for more general trajectories. The main potential application of the presented method is detection of abnormal behavior of humans, hand gesture detection and recognition, etc. Hidden Markov Models (HMMs) and Gaussian Mixture Models are used as the base of the the classification mechanism. The paper contains description of the method, description of experiments and their results, and conclusions.

Keywords

Trajectory classification, Gaussian Mixture Models, Hidden Markov models, video sequences.

1.INTRODUCTION

Understanding dynamic behavior of objects is still very open problem. One task, which is a small subset in the wide objects dynamics field, is classification of short fragments of motion of objects. This task is particularly interesting for human computer interfaces, computer applications in surveillance tasks, computer-assisted communication between humans, general computer graphics and computer vision, etc. Human computer interfaces can benefit specifically from understanding human hand gestures and human body motion. Surveillance applications can exploit detection of “abnormal” behavior of humans in e.g. public places. Computer-assisted communication between humans can benefit from understanding the gestures and body motion, too, specifically in case the users cannot see or even hear each other or in cases where the communication is otherwise limited. Other applications in computer graphics and computer vision include e.g. human motion understanding in navigation in virtual 3D space, animation of objects and humans, or motion capture assistance.

Previous Work

The generally used approach to analysis of dynamic behavior of objects is focused on understanding of video content. This leads to decomposition of problem to a couple levels of analysis. The first level of decomposition usually consist of image processing e.g. object recognition. This part also includes temporal element e.g. object tracking. The problem of object recognition and tracking is discovered in [Hra06], [Jir06], [Jir08]. The second level of decomposition is information interpretation. Decomposition of problem and solution of parts is widely discussed in [Mli08]. The [AC99], [CRCZ05] summaries the approach with focus on human motion analysis.

More general low level video processing which tends to obtaining complex information from scene could be based for example on texture analysis [CV05], important points analysis (SIFT) or areas analysis (SURF) [Bay08]. However, it is very difficult to describe behavior patterns in general, therefore proposed approach use only informations about well defined objects in scene.

The result of content focused analysis is set of features with describes variable count of objects in time. The special case of these features in time is the trajectory. For detailed analysis of object behavior is essential to describe all possible behaviors (trajectories) and assign them the most appropriate interpretation, which is discussed in [Mli08a]. In some cases, few well defined “strong types of behavior” could be sufficient for scene understanding.

The Gaussian Mixture Models (GMM) method is suitable for the trajectory classification task as it is capable of recognition of pattern in temporal variable data. The GMM statistically describes whole feature vector in time. Other well known approach, Dynamic Time Wrapping (DTW), describes changes of feature vector in time. The generalization of these two methods is done by Hidden Markov Models (HMM) which was used for sign language recognition [Sta95] and gesture recognition [Rig97], [Bor08].

The method of gesture classification through the alternative Conditional Random Fields is discussed in [WLD06].

Data and Trajectory Classes

Based on the results of the previous work, we examined some data and methods for dynamic gesture processing and human motion trajectories. The data was available from CareTaker [Car08] in the form of moving humans in a public environment (underground stations) and AMIDA [Ami08] projects in the form of meeting data (humans sitting by a table and having a meeting). Since the number of all possible classes in data and their inner-class variability is too high, we have chosen and limited our solution to only few classes of trajectories. These classes included so-called Speech Supporting Gestures (SSG) and Abnormal Human Trajectories (AHT). The choice was made with the motivation of improvement of the speaker identification in the meeting data and of identification of generally abnormal behavior of humans in the public areas.

Our approach can be subdivided into several parts. The first part of the approach was to obtain the object trajectories through video sequence processing. Next part is dynamic (pre)processing of the trajectories. Finally, the training and classification is the last part of the processing chain.

2.OBJECT TRAJECTORIES

The aim of this part is to find positions and sizes of the objects that we can affirm to correspond to body parts. This approach has been widely investigated as mentioned above but let us briefly mention the approaches used in the experiments.

Object Localization

The localization task starts with the color-based segmentation. A Single Gaussian Method trained from two sets (skin and non-skin pixel colors) outputs the Skin Probability Image (SPI) where connected components are possible occurrences of ROIs. Every region area is evaluated and only the ones with the largest size are considered.

Once these regions are found we need to assign them to the body parts presence of which is expected in the image. For the available meeting data, simplified approach has been taken so the number of persons in every meeting data sequence is assumed to be two as well as it is assumed that in the beginning of each one, both persons remain in regular positions (head above hands and hands not switched while each person occupies his/her “half” of the image). In the public environment data, every moving object is considered to be a moving human. This approach has proven to be sufficient for the further presented steps. (See Figure 1 for the meeting data approach.)

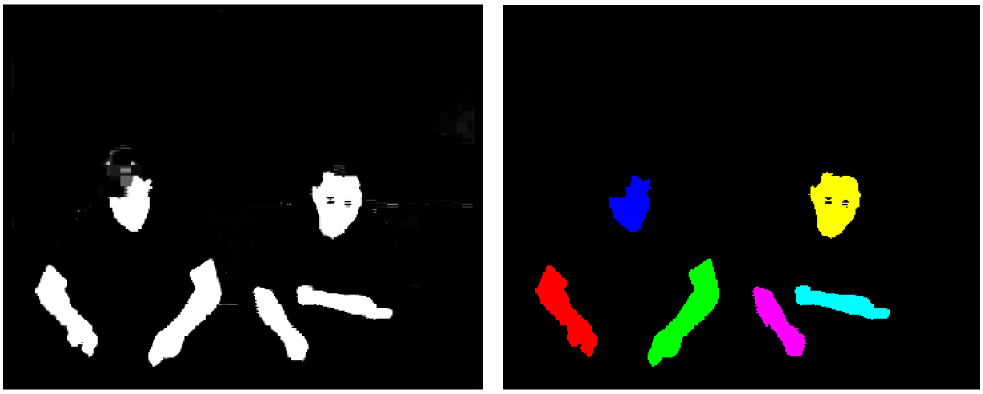


Figure 1. Skin probability and connected components evaluation

Object Detection

The positions of participants’ hands can be refined by localizing palms instead of hands (entire skin-colored regions). Let us assume the sub-image containing hand which has been multiplied by appropriate SPI. In such a sub-image only a flat region of skin color will occur corresponding to the elbow (and potentially arm) and a region with a certain number of protrusions as a consequence of inter-finger distance and shadows. Then, local maxima clusters feature finger region in this sub-image convolved with kernels. (See Figure 2 for illustration of the approach.)

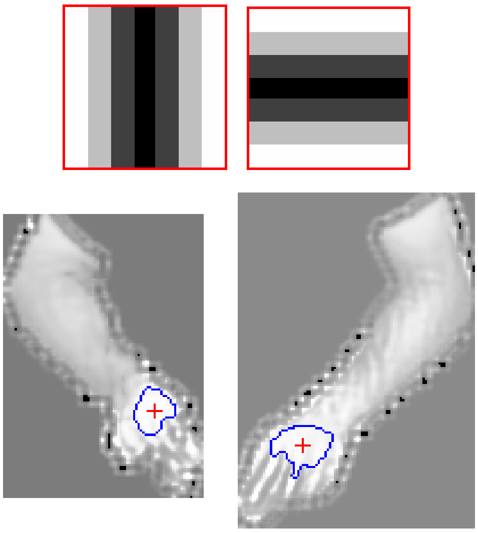


Figure 2. Convolution kernels and hand localization

Object Tracking

Once the regions are identified, they are consequently tracked so their trajectories are found.

From all of the inquired algorithms for object tracking, the Overlapping Boxes Method [Wah07] was chosen as it showed the best results on the meeting data. The essence of this method is that in two subsequent images, the two occurrences of the same object lie so close to each other that the bounding boxes overlap.

However, this algorithm cannot solve the occlusion problem. The problem of occlusion could be solved by modeling of motion by KLT filter presented e.g. in [Hra06]. In the future work, more tracking algorithms should be investigated; however, for the experiments shown below, the simple tracking algorithms are sufficient.

In case of public areas, the input data in the form of positions of humans were available along with the data set so the positions of the humans in the video were known in advance and no need for application of tracking algorithms existed.

3.TRAJECTORY PREPROCESSING

The trajectory can be described as a timed sequence of tuples [*x*, *y*] *x*, and *y* represent positions of the object. For the classification purposes, it is not clear, whether the motion of objects is characterized by its absolute position, speed of changes of the position (first derivative of position), or perhaps even acceleration (second derivative of the position). Additionally, co-ordinate system might be of interest in this context. In the presented approach, the quadruple [*x*, *y*, *dx*, *dy*], where *x*,*y* represents positions of object and *dx*, *dy* represent its velocity, was chosen as a good compromise.

Classification process selects trajectories (represented here through a sequence of the above mentioned quadruples) that do belong to a certain class (or classes). However, as we need to classify only parts of trajectories that are characteristic for the class (classes) of interest. A problem can occur in situation where the trajectory is relatively long and/or in cases where the characteristic motion is preceded or bound with other motion. This fact can lead in a need to cut the trajectories.

The method of choosing the best position of cut is unknown in general, but it depends on the application. The usual way is to divide the trajectory on uniform overlapping parts. Another approach assumes its division according to quiescent state of trajectory (trajectory has small energy).

By examining the trajectories a certain degree of jitter was discovered. This might cause unsatisfactory result in the recognition stage. So the Double Exponential Filtering was worked in. This filter makes the positions of tracked objects stable when in a rest phase preventing unwanted sub-trajectories from being segmented out by the Activity Measure Method (AM). In a dynamic phase (hand movement, etc.) it ensures that a motion trend from previous images is taken into account resulting in smoother trajectory.

In case, when we want to process position coordinates of the trajectory, it is necessary to normalize it. The normalization process consists of alignment of all coordinates according to mean position.

4.TRAINING AND CLASSIFICATION

As mentioned in the previous part, the purpose of classification is to decide whether a trajectory or a sub-trajectory belong to a certain class. Such a sub-trajectory may represent e.g. gesture.

For the experimental purposes, a set of trajectories was manually segmented and modeled using one Gaussian Mixture Model M for each of the two basic classes we had decided to work with: one for those led in horizontal direction (M_H) and one for those in vertical (M_V). The basic models of this situation are shown in Figure 3.

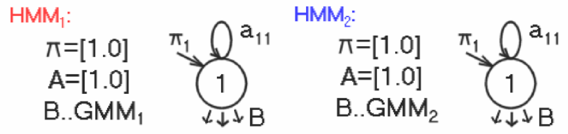


Figure 3. Hidden Markov Model corresponding with single Gaussian Mixture Model

The Figure describes Hidden Markov Model with the unitary transition probabilities and the emitting probability given by Gaussian Mixture Model.

For every unknown potentially infinite sequence $O(n) = (o_1, o_2 .. o_n)$ the log-likelihood of being emitted by the horizontal or vertical model can be computed as follows:

$$p(O|M_x)=\sum_{i=1}^{n}\frac{-\log p(o_i|M_x)}{n}$$

As an auxiliary metrics for validation of O belonging in SSG class we have defined periodicity of O in this manner: suppose some subsidiary vector $w = (w_1, w_2 .. w_Z)$ elements of which (w_i) are indices of winning distributions in model M_X for o_i where M_X stands for the model with higher $p(O | M_X)$. The periodicity is then the number of sub-sequences ($w_i, w_{i+1} .. w_j$) for which the following conditions are met:

$i \geq 1, j \leq Z, j - i \geq C$

$$\forall k=i..(i-1): w_k=w_{k+1}$$

$w_k = w_{k+1}$ where $k = i .. j - 1$

$w_{i-1} \neq w_i$

$w_{j+1} \neq w_j$

In the first condition we can find a constant C which defines the minimal length of such a sub-sequence (period).

As it has been observed that the higher the number of periods is the more probably the unknown sequence will be a SSG class representative.

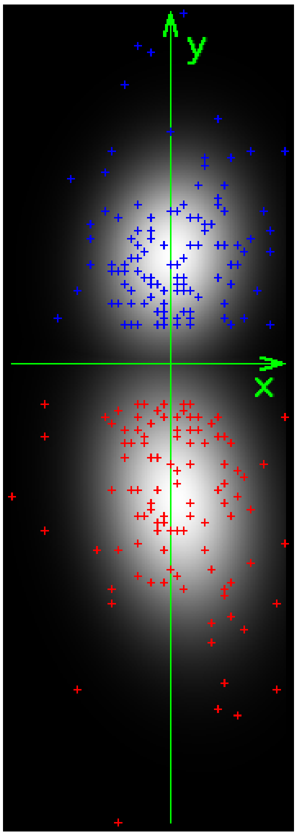
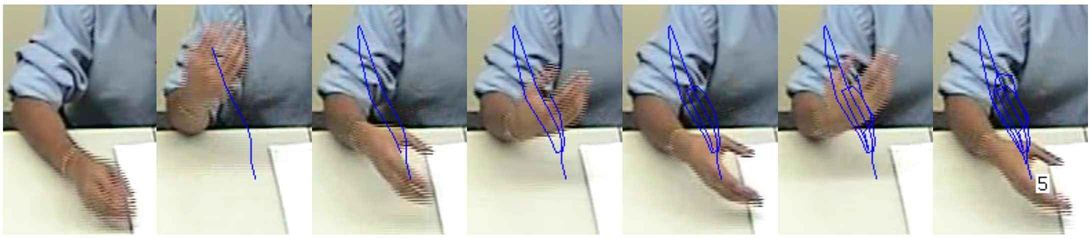


Figure 4. Gaussian Mixture Model for “vertical gestures” sub-class

The Gaussian Mixture Models describing two SSG and non-SSG are shown in Figure 4. The red colored cluster (in bottom part of picture) denotes non-SSG clusters and blue cluster denotes SSG gestures.



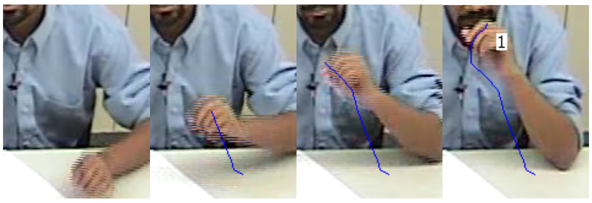


Figure 5. Examples of Speech Supporting Gesture and no-gesture

The training and evaluation of speech supporting gestures was done on data set containing about 30 trajectories with average length of 20 coordinates. The accuracy of classification was about 62%.

However, the single state HMM cannot differ more complex gestures. The second experiment was done with data from a public environment. For single scene was defined five simple classes which denoted different person behavior. For example person going thru turnpike (e.g. entering platform) and person going around turnpike.

The classification algorithm for Hidden Markov Model M is realized by the Viterbi Algorithm [Cer03] and it is defined as finding of path thru model M with best response:

$$P^*(O|M)=\max_X P(O,X|M)$$

The probability of passing of Object O thru a model M by way X is described by following equation, whereas $a_{k,j}$ denotes transition probabilities between states. Except first and last state, states are emiting or generating output probability density function $b_j(o(t))$.

$$P(O,X|M)=a_{x(o)x(1)}\prod_{n=1}^{T} b_{x(t)}(o_t)a_{x(t)x(t+1)}$$

For this case it was chosen more complex HMMs and the topology of each of them is shown in Figure 5.

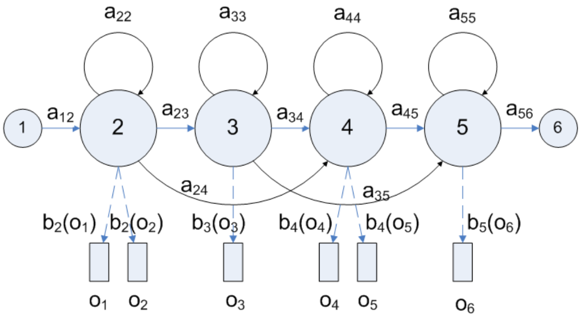


Figure 5. Example of non trivial Hidden Markov Model

Training and evaluation was performed on about 800 trajectories. The trajectory correct class was recognized with accuracy about 90% on the given data.

5.CONCLUSIONS

The work presented in this paper focused on classification of hand gestures and human motion. It has been shown that the classification of the trajectories through the HMM classifier is feasible and can lead into good results.

The single state HMM is well for recognition of gesture parts or simple gestures with single direction. Opposite, more complex gestures are very interesting from point of human behavior analysis and understanding. However, definition and training of complex gestures is more difficult.

However, we believe that the presented approach is reasonably well working and generally usable.

Further work is needed specifically in the field of refinement the sub-trajectories selection and also in refinement of representation of the trajectories.

The performance of the presented algorithm is affected by the segmentation and tracking methods. And the best working methods to be used in the presented approach have yet to be evaluated.

The experiments has shown, that the HMM based approach is suitable for sequential data like are gesture trajectories.

6.ACKNOWLEDGMENTS

This work has been supported by AMIDA - Augmented Multi-party Interaction with Distance Access, EU-6FP-IST, IST-033812-AMIDA, by the Ministry of Education, Youth and Sports of the Czech Republic under the research program LC-06008 (Center for Computer Graphics), and by the research project Security-Oriented Research in Information Technology, MSM0021630528.

7.REFERENCES

[AC99] Aggarwal, J. K., Cai, Q.: Human motion analysis: a review.Comput. Vis. Image Underst., 73(3):428–440, 1999.

[Ami08] Web site. Project: Augmented Multi-party Interaction with Distance Access (AMIDA), 2008. http://www.amiproject.org/

[Bay08] Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. 2008. Speeded-Up Robust Features (SURF). *Comput. Vis. Image Underst.* 110, 3 (Jun. 2008), 346-359.

[Bis06] Bishop, C. M. 2006. Pattern Recognition and Machine Learning(Information Science and Statistics). Springer-Verlag New York,Inc., Secaucus, NJ, USA.

[Bor08] Borza P.-V.: Motion-based Gesture Recognitionwith an Accelerometer, Bachelor's Thesis in Computer Science, Babeş-Bolyai University of Cluj-Napoca, Romania, 2008

[Car08] Web site. Caretaker. Information Society Technologies, 2008. http://www.ist-caretaker.org/.

[Cer03] Černocký, J. 2003. Hidden markov models – an introduction.Tech. rep., Brno University of Technology, Faculty of Information Technology.

[CV05] Antoni B. Chan and Nuno Vasconcelos. Mixtures of dynamic textures. In ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, pages 641–647, Washington, DC, USA, 2005. IEEE Computer Society.

[Has06] N. Hassink and M.G. Schopman, "Gesture recognition in a meeting environment," Feb. 2006; http://essay.utwente.nl/56211/.

[Hra06] Hradiš Michal, Juránek Roman: Real-time Tracking of Participants in Meeting Video, In: Proceedings of The 10th Central European Seminar on Computer Graphics, Budměřice, SK, 2006, p. 5

[Jir08] Jiřík, L.: Recognition of poses and gestures. Master's thesis, Brno University of Technology, Faculty of Information Technology, 2008.

[Jir06] Jiřík, L.: Detection Human Body Parts. Bachelor's thesis, Brno University of Technology, Faculty of Information Technology, 2004.

[Mli08a] Mlích, J., Chmelař, P.: Trajectory classification based on hidden markov models. In Proceedings of 18th International Conference on Computer Graphics and Vision, pages 101–105. Lomonosov Moscow State University, 2008.

[Mli08] Mlích, J:. Object tracking in video sequences. Master's thesis, Brno University of Technology, Faculty of Information Technology, 2008.

[Nam96] Y. Nam and K. Wohn, "Recognition of space-time hand-gestures using hidden Markov model," ACM Symposium on Virtual Reality Software and Technology,pp. 51–58, 1996.

[Sta95] Starner, T.: Visual recognition of American Sign Languageusing hidden Markov models. Master's thesis, 1995 .

[Rig97] G. Rigoll, A. Kosmala, and S. Eickeler, "High Performance Real-Time Gesture Recognition Using Hidden Markov Models," In Proc. Gesture Workshop,, pp. 69--80, 1997.

[Row97] Roweis, S., Ghahramani, Z.:A unifying reviewof linear Gaussian models. Tech. rep., 6 , King's College Road,Toronto M5S 3H5, Canada, 1997.

[Sta00] Stauffer, C., Grimson, et al.: Learning patterns of activity using real-time tracking. IEEE Transactions onPattern Analysis and Machine Intelligence 22, 8, 747–757. 2000.

[You06] Young, S., Odell, et al.: The HTK Book, University Of Cambridge, UK, 2006

[WLD06] Wang S.B, Quattoni A., et. at.: Hidden Conditional Random Fields for Gesture Recognition, Computer Science and Artificial Intelligence Laboratory, MIT , 2006.

[Wah07] Wahde, M. a kol. Computer Vision Based System for Dynamic Gesture Recognition. Göteborg, Technical Report, Sweden, 2007.