Gesture Segmentation from a Video Sequence Using Greedy Similarity Measure

Qiulei Dong, Yihong Wu and Zhanyi Hu
National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, P.O. Box 2728, Beijng 100080, P.R. China
{qldong, yhwu, huzy}@nlpr.ia.ac.cn

Abstract

We propose a novel method of greedy similarity measure to segment long spatial-temporal video sequences. Firstly, a principal curve of motion region along frames of a video sequence is constructed to represent trajectory. Then from the constructed principal curves of trajectories of predefined gestures, HMMs are applied to modeling them. For a long input video sequence, greedy similarity measure is established to automatically segment it into gestures along with gesture recognition, where true breakpoints of its principal curve are found by maximizing the joint probability of two successive candidate segments conditioned on the gesture models obtained from HMMs. The method is flexible, of high accuracy, and robust to noise due to the exploitation of principal curves, the combination of two successive candidate segments, and the simultaneous recognition. Experiments including comparison with two established methods demonstrate the effectiveness of the proposed method.

1. Introduction

Video segmentation is to decompose a long video sequence into a number of short representative video segments. It has many prospective applications in video indexing, video searching, visual surveillance, etc.

In many previous methods [1, 3, 7], shot transition detection was used frequently for video segmentation. For example, Porter et al. [7] proposed a unified method to detect shot transitions including cuts, fades and dissolves. This kind of segmentation approach based on shot transition detection is not related to the video content explicitly, but just to find the end of a continuous video shot which may consist of several distinct video segments. What's more, if there is only one shot in video sequence, this kind of segmentation method may be futile.

Therefore, in order to do segmentation more accurately, many segmentation methods [6, 8] based on motion content have been adopted in recent years. Peyrard and Bouthemy [6] proposed an automatic content-based segmentation approach, which relies on the analysis of the temporal evolution of the dynamic video content. In addition, Xiang and Gong [11] proposed an on-line segmentation algorithm to detect breakpoints in the video sequence. In [5, 10], motion trajectories were used for temporal segmentation of activities from continuous video sequence. The trajectory was usually represented using the curve obtained by fitting representative points of motion regions or using the polygonal line obtained by connecting representative points sequentially. However, this kind of representation method is based on the assumption that these points submit to a certain distribution, which inevitably limits its application.

In this paper, based on gestures, we propose a novel video segmentation approach using greedy similarity measure. The main steps and contributions are as follows:

1. Trajectory representation. We use a principal curve [2, 9] to describe the trajectory of the motion region along frames in 3-d space. The principal curve is constructed by using the 'soft' version of k-segments algorithm in [9] and is improved by merging points at temporal intermission. It adapts well for representing not only general but also complex trajectories.

2. Gesture modeling. From the constructed principal curves of trajectories of predefined gestures, HMMs are applied to modeling them.

3. Video segmentation into gestures. For an input unknown long video sequence, we establish a greedy similarity measure to automatically detect the true breakpoints on the constructed principal curve from its trajectory. Firstly, principal curve is constructed and then the point set at each intermission is merged into one point. The mergence makes the segmentation more accurate. The true breakpoints are obtained by minimizing the proposed greedy similarity measure function of two successive candidate segments, where the function is on the probabilities conditioned on the gesture models given by HMMs. Thereafter, the long

0-7695-2521-0/06/$20.00 (c) 2006 IEEE

trajectory is segmented into shorter segments corresponding to different recognized gestures.

Experiments on hip-hop dance videos are performed. And then the results are evaluated by $Recall$ and $Precision$, two evaluation criterions, widely used in many segmentation methods [1, 7]. The evaluation shows that our method can achieve better results compared with the DCE method [4, 11] and the HMM-based method [5].

The remainder is organized as follows. Section 2 describes the trajectory representation based on principal curve and trajectory merging. Section 3 gives the gesture modeling by HMMs. The proposed greedy similarity measure segmentation is presented in Section 4. Experiments are reported in Section 5. And some concluding remarks are listed in Section 6.

2. Trajectory representation and merging

2.1. Trajectory representation

Usually, motion trajectory is represented using curve obtained by fitting representative points of motion regions or using polygonal line obtained by connecting representative points sequentially along frames of video sequence. This kind of representation is based on the assumption that these points submit to a certain distribution. However, this assumption may be unreasonable in many situations.

In order to ensure that trajectory analysis is not based on a wrong assumption, we introduce a trajectory representation using principal curve. A principal curve [2, 9] describes the intrinsic structure of a data set based on nonparametric analysis that is shown suitable for representing a trajectory.

There are two steps to obtain the trajectory representation in our representation method. First, the point set of the motion regions along time axis is obtained by tracking sequentially. Then, the 'soft' version of k-segments algorithm [9] is used for constructing the principal curve from the point set in 3-d space(Figure 3).

2.2. Merging algorithm

For an input unknown long video, its constructed principal curve is improved as follows:

If there is a few frames' intermission between two successive gestures, the corresponding curve segment on the principal curve from these frames, denoted by cs, is roughly perpendicular to the image planes of these frames, which means that the points (x_i, y_i, t_i) on cs are subject to:

$$\max_i \sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2} \leq \delta \qquad (1)$$

where δ is a predefined threshold, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, n is the total number of the points on cs.

The intermission frames, if their number is less than a predefined threshold, will be ignored. So, we also consider another constraint:

$$Numf \geq N_0 \qquad (2)$$

where $Numf$ is the number of intermission frames, N_0 is a predefined threshold.

Satisfying the above two constraints (1) and (2), the points on the principal curve segment are merged into one point. This step makes the representation more accurate.

3. Gesture modeling

HMMs have been proved to be able to model activities characterized by spatial-temporal structures efficiently. So we use HMMs to model different basic gestures in this work.

From an input basic gesture, its principal curve is constructed. Then the principal curve is modeled by HMM, where Gaussian mixture distribution is assumed as the probability distribution for each state(due to the space limit, the detail of HMMs is omitted). Let T_{min} be the minimal time duration, T_{max} be the maximal time duration from all the gestures, N be the number of gesture models, and $M_i (i = 1, ..., N)$ be the i-th gesture model.

4. Video segmentation into gestures

Denote the principal curve of an input unknown long video as PC and then let s and e be any two points on it. The curve segment from PC between s and e is denoted as $C(s, e)$. The probability for $C(s, e)$ conditioned on M_i being the true gesture model is denoted as $P(C(s, e)|M_i)$.

Based on the following two conditions, a point e is considered as a candidate breakpoint and a model M_i as the corresponding candidate gesture model:

c1) $logP(C(s, e)|M_i)$ is the maximum in the set of all gesture models on the point e, and is the local maximum in the neighborhood of e on PC simultaneously.

c2) Let $logP(C(s, e)|M_j)$ be the second maximum in the set of all gesture models on the point e, the difference between $logP(C(s, e)|M_i)$ and $logP(C(s, e)|M_j)$ is larger than a predefined threshold.

Then the proposed segmentation algorithm is given as follows: The input for this segmentation procedure is the principal curve PC, and the start point of the first segmentation is chosen at the first point on PC. This segmentation procedure is iterative. Assume we have obtained $(k - 1)$ segments along PC and the start point of the k-th segmentation on PC. Let this start point be s_k.

Starting from s_k along PC after T_{min} but within T_{max} we search a point e_k, and in the set of all gesture models we search a gesture model M_k, with (s_k, e_k, M_k) satisfying the above two conditions c1) and c2). The number of the searching results for (e_k, M_k) is generally not unique, and we denote the resulting set as $CS_k = \{(e_{kq}, M_{kq}), q = 1, ..., n_k\}$ indexed by q. For a (e_{kq}, M_{kq}) in CS_k, e_{kq} is a candidate breakpoint for the k-th segmentation and M_{kq} is the corresponding candidate gesture model. For each e_{kq} from CS_k, we search the farthest point $(e_{kq} + \lambda_{kq})$ along PC satisfying:

$$d(e_{kq}, e_{kq} + \lambda_{kq}) < \varepsilon \tag{3}$$

where $d(e_{kq}, e_{kq} + \lambda_{kq})$ is the Euclidean distance in 3-d space, and ε is a predefined threshold. The point $(e_{kq}+\lambda_{kq})$ is set as the start point of the $(k+1)$-th temporary segmentation. Similarly starting from $(e_{kq} + \lambda_{kq})$ and repeat the above process, we obtain the candidate breakpoints of the $(k+1)$-th temporary segmentation. The set of these candidates is denoted as $CS_{k+1} = \{(e_{(k+1)p}, M_{(k+1)p}), p = 1, ..., n_{k+1}\}$ indexed by p.

Then a greedy similarity measure function is constructed as:

$$\begin{aligned} &\xi(e_{kq}, M_{kq}, \lambda_{kq}, e_{(k+1)p}, M_{(k+1)p}) \\ &= -log(P(C(s_k, e_{kq})|M_{kq}) \\ &\quad \times P(C(e_{kq} + \lambda_{kq}, e_{(k+1)p})|M_{(k+1)p})) \\ &= -logP(C(s_k, e_{kq})|M_{kq}) \\ &\quad - logP(C(e_{kq} + \lambda_{kq}, e_{(k+1)p})|M_{(k+1)p}) \end{aligned} \tag{4}$$

where (e_{kq}, M_{kq}) varies in CS_k, $(e_{(k+1)p}, M_{(k+1)p})$ varies in CS_{k+1}. The corresponding (e_{kq}, M_{kq}) minimizing (4) are considered as the true breakpoint and the corresponding recognition of the k-th segmented gesture respectively. Then the point $(e_{kq} + \lambda_{kq})$ is set as the start point s_{k+1} of the $(k+1)$-th segmentation.

Starting from s_{k+1}, repeat this process until the length of the remaining frames is less than T_{min}. Finally, an input unknown long video is segmented into different shorter recognized gestures. A flowchart of the segmentation algorithm is shown in Figure 1.

This segmentation method is called greedy similarity measure (GSM). The reasons we call it as GSM are:

1. In function (4), each probability term represents the similarity between the corresponding curve segment and gesture models.

2. It looks for all the candidate breakpoints during two successive curve segments iteratively. And the true breakpoint is identified by minimizing the function (4).

It is obvious that along with our segmentation, each segmented result is also recognized as a gesture simultaneously.

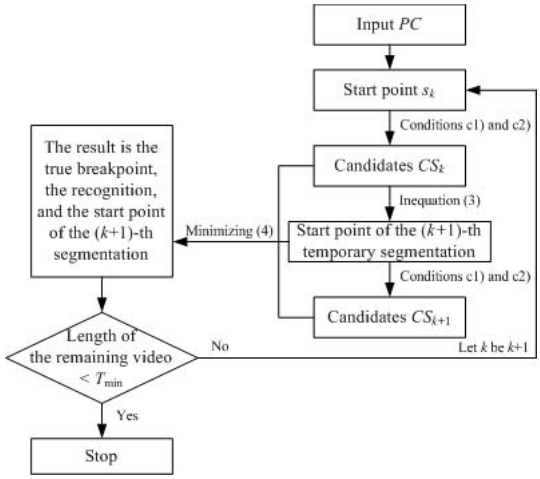

Figure 1. Flowchart of the proposed segmentation algorithm.

5. Experiments

Due to the space limitation, only the experiments on hip-hop dance are shown here and other experiments having similar results are omitted. The used sequences, each of which contains twelve gestures, are performed by eight different people for eight different times. We have 64 sequences in total. These sequences are captured by a digital camcorder and each of them has 300-700 frames. The resolution of each frame is 300×240 pixels. Figure 2 shows two gestures from a hip-hop sequence.

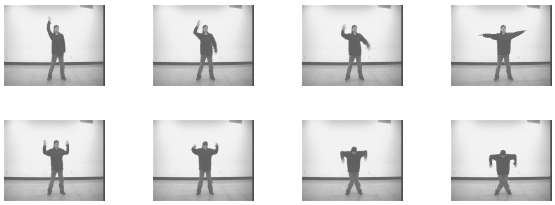

Figure 2. Two predefined gestures in our experiments. Each row corresponds to one predefined gesture.

We arbitrarily select 32 sequences performed by four people for training to obtain gesture models with manual segmentation. The remaining sequences are used for testing. The main motion regions in the hip-hop sequences are from the moving hands. So the principal curves are constructed from the moving hands of the dancers and then the point set at each intermission is merged into one point. A principal curve representing the trajectory of a hand is shown in Figure 3(a).

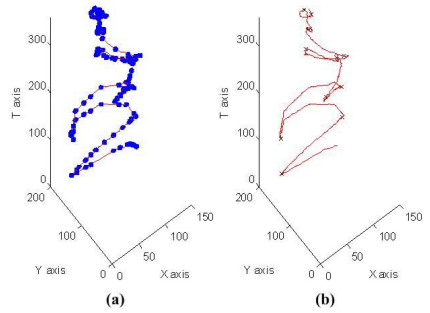

Figure 3. (a) A principal curve passing through the motion regions. (b) An example of the detection results.

Applying the proposed algorithm of the greedy similarity measure, those remaining 32 sequences are segmented automatically. The final breakpoint detection results are shown in the column of GSM in Table 1. And an example of the detection results is shown in Figure 3(b).

The detection results are evaluated by two measures, $Recall$ and $Precision$:

$$Recall = \frac{N_{correct}}{N_{correct} + N_{missed}} \quad (5)$$

$$Precision = \frac{N_{correct}}{N_{correct} + N_{false}} \quad (6)$$

where $N_{correct}$ is the number of correctly detected breakpoints, N_{missed} is the number of missed breakpoints, and N_{false} is the number of falsely detected breakpoints. $Recall$ and $Precision$ are usually used as two evaluation criterions in video segmentation. It is obvious that the closer the values of $Recall$ and $Precision$ are to 1, the more effective the method is. In our detection results, the value of $Recall$ is 93.2% and the value of $Precision$ is 91.6% both shown in Table 1.

In addition, we compare the proposed method with the DCE method [4, 11] and the segmentation method based on HMMs [5]. The performances of all the three methods are presented in Table 1.

Table 1. Experimental results

	DCE	HMM	GSM
True breakpoints	384	384	384
Correctly detected	247	275	358
Falsely detected	91	119	33
Recall	64.3%	71.6%	93.2%
Precision	73.1%	69.8%	91.6%

6. Conclusions

A novel method for gesture segmentation from a video sequence using greedy similarity measure is proposed and validated by hip-hop sequences. The main characteristics of the proposed method are: 1) Principal curve describes more faithfully the intrinsic structure of motion regions based on nonparametric analysis. 2) Greedy similarity measure is flexible and more robust to noise due to the exploitation of the combined information from two successive gestures. 3) The segmentation and gesture recognition are carried out simultaneously, which makes the result more accurate.

In future, our method will be extended to other applications such as behavior segmentation and video indexing.

Acknowledgment

This work was supported by the National High Technology R&D Program of China under the grant No. 2005AA114130.

References

[1] J. S. Boreczky and L. A. Rowe. Comparison of video shot boundary detection techniques. In *Proc.SPIE Conf. SRIVD*, pages 170–179, San Jose, 1996.

[2] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84(406):502–516, 1989.

[3] J. Kender and B. Yeo. Video scene segmentation via continuous video coherence. In *Proc. IEEE Conf. CVPR*, pages 367–373, Santa Barbara, CA, Jun. 1998.

[4] L. J. Latecki and R. Lakämper. Convexity rule for shape decomposition based on discrete contour evolution. *Comput. Vis. Image Underst.*, 73(3):441–454, 1999.

[5] J. Min and R. Kasturi. Extraction and temporal segmentation of multiple motion trajectories in human motion. In *Proc. IEEE Workshop on Detection and Recognition of Events in Video*, Washington DC, 2004.

[6] N. Peyrard and P. Bouthemy. Content-based video segmentation using statistical motion models. In *Proc. BMVC*, pages 527–536, Cardiff, 2002.

[7] S. V. Porter, M. Mirmehdi, and B. T. Thomas. Temporal video segmentation and classification of edit effects. *Image and Vision Computing*, 21(13-14):1097–1106, Dec. 2003.

[8] Y. Rui and P. Anandan. Segmenting visual actions based on spatio-temporal motion patterns. In *Proc. IEEE Conf. CVPR*, pages 111–118, Hilton Head, SC, 2000.

[9] J. Verbeek, N. Vlassis, and B. Kröse. A k-segments algorithm for finding principal curves. *University of Amsterdam, The Netherlands, Technical report, IAS-UVA-00-11*, 2000.

[10] F. Wang, C. Ngo, and T. Pong. Gesture tracking and recognition for lecture videoediting. In *Proc. ICPR*, pages 934–937, Cambridge, UK, Aug. 2004.

[11] T. Xiang and S. Gong. Activity based video content trajectory representation and segmentation. In *Proc. BMVC*, pages 177–186, Kingston, 2004.