Multi-layer Spectral Clustering for Video Segmentation

Xiaofei Di[1,2], Hong Chang[1], Xilin Chen[1]

[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China
[2]Graduate School of the Chinese Academy of Sciences, Beijing, 100039, China
{xiaofei.di, hong.chang, xilin.chen}@vipl.ict.ac.cn

Abstract. Video segmentation with spatial priority suffers from incoherence problem, since the presegments of consecutive frames may be very different. To address this problem, this paper proposes an effective and scalable approach for video segmentation, aiming to cluster video pixels that are coherent in both appearance and motion. We build up a multi-layer graph based on multiple segmentations of the video frames, where each presegment corresponds to a vertex in the graph and each layer corresponds to the segmentation result using mean shift algorithm under specific granularity. Three types of edges are connected in the graph and the corresponding affinities are defined which convey local grouping cues of intra-frame, inter-frame and inter-layer neighborhoods. Then the task of video segmentation is formulated into graph partition, which can be solved efficiently by power iteration clustering algorithm. Both qualitative and quantitative experimental results demonstrate the efficacy of our proposed method.

1 Introduction

Segmentation is an important research topic in computer vision. Video Segmentation generalizes the concept of Image Segmentation to spatio-temporal space and aims to cluster pixels, which are coherent in both appearance and motion, into volumes. It is fundamental for many high-level computer vision tasks such as region-based video coding, object tracking, activity recognition, motion analysis, 3D scene analysis, and content-based retrieval.

Existing video segmentation methods can be classified into three categories. 1) video segmentation with temporal priority (a.k.a. motion segmentation) [1][2] groups salient image feature trajectories, and then gets the label of every pixel from these features. However, getting perfect feature trajectories relies on feature tracking which is also a hard problem. 2) simultaneous spatial-temporal video segmentation in x-y-t space. Algorithms of this class are always generalized from image segmentation and can be divided into either non-graph-based [3][4] or graph-based [5][6][7]. All of them directly deal with the large number of pixels in the video, which is costly in both memory and computation time.

3) video segmentation with spatial priority (a.k.a. layer extraction)[8][9][10] segments each frame at first, and matches these presegments across all frames. During the past decades, image segmentation techniques have been widely studied and made great progress. Therefore it is prosperous to segment a video based on the presegments of all frames, which is also the focus of this paper.

Unfortunately, for the real world images of the same scene, when captured under slightly different conditions, their segmentation results could be very different. The unstable (or inconsistent) segmentation also exists for consecutive video frames, leading to incoherent presegments. The second row of Fig.1 illustrates this problem of mean shift segmentation on consecutive video frames. Many image segmentation methods, besides mean shift, suffer from this issue, which makes spatial prior video segmentation a big challenge. Several previous works have tried to solve this problem. [9] maps and clusters homography matrices between these presegments in a low dimensional linear subspace where it is proven that they form well-defined clusters. [8] finds the optimal corresponding region pairs through circular dynamic-time warping (CDTW), where the motion properties are captured by homography matrices. Whereas, [10] directly concerns about the region mismatch problem and matches presegments according to partial matching cost which is based on affine transformations. All above methods model the relationship between presegments in parametric ways, i.e. homography matrix and affine transformation. Nevertheless, the inconsistency problem results in very diverse regions whose relationship cannot be perfectly captured using parametric models.

As shown in Fig.1, the coarser the segmentations, the more stable the results. Inspired by the empirical experiments as well as several recent multi-layer image segmentation approaches [11], we propose a new multi-layer spectral clustering algorithm for video segmentation in this paper. The main contributions lie in the following two aspects:

1. We present a non-parametric graph-based video segmentation method with both spatial and temporal consistency. More specifically, we construct a multi-layer graph over the initial segmentation results using mean shift algorithm with different granularities, where the inter-layer affinities can effectively combine coarse-but-stable and fine-but-unstable details, and the affinities between consecutive frames can provide local grouping cues across temporal range. Note that [11] constructs a multi-layer graph similarly by varying the parameters of the mean shift algorithm. The main differences between [11] and our method include: [11] focuses on integrating local grouping cue in image segmentation, while ours focuses on high-quality video segmentation with spatial and temporal consistency; in [11], both pixels and presegments (regions) are considered as graph nodes while ours only considers presegments, and the affinity measures are defined differently as well.
2. Our proposed method is efficient and scalable. We adopt spectral segmentation framework [12][13][14] in this paper. Once obtained the affinities of the multi-layer graph, we perform power iteration clustering [15], which is fairly efficient, to get the final video segmentation. Moreover, we propose a

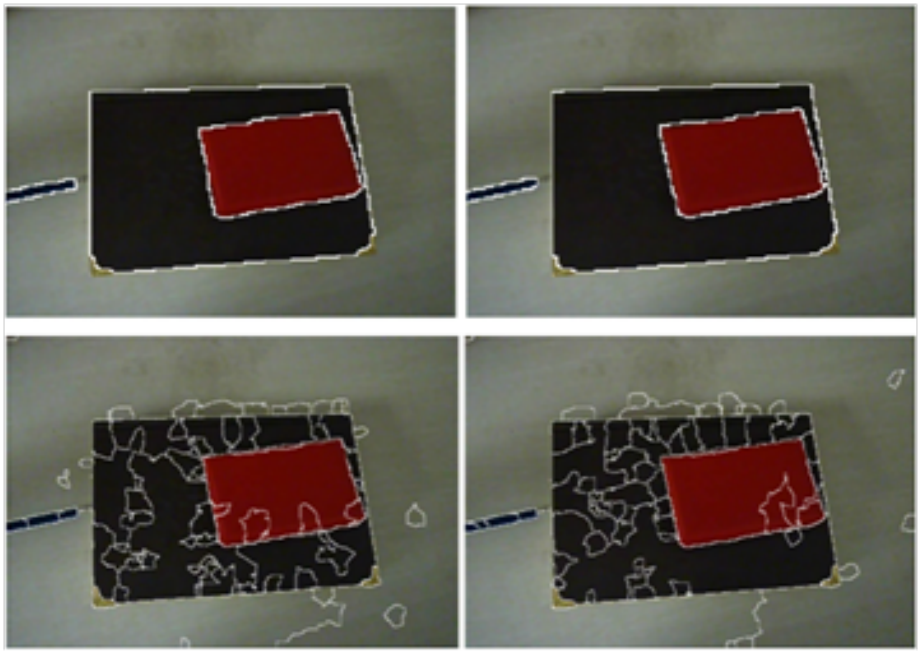


Fig. 1. Mean shift segmentation on two consecutive video frames under different granularities (through parameter tuning). Top row: coarser segmentations. Bottom: finer segmentations.

scale-up strategy to make our method feasible for high-resolution and long video clips.

To summarize, our proposed method is effective, efficient, and scalable for real-world video segmentation. Both qualitative and quantitative experiments verify the efficacy of our algorithm.

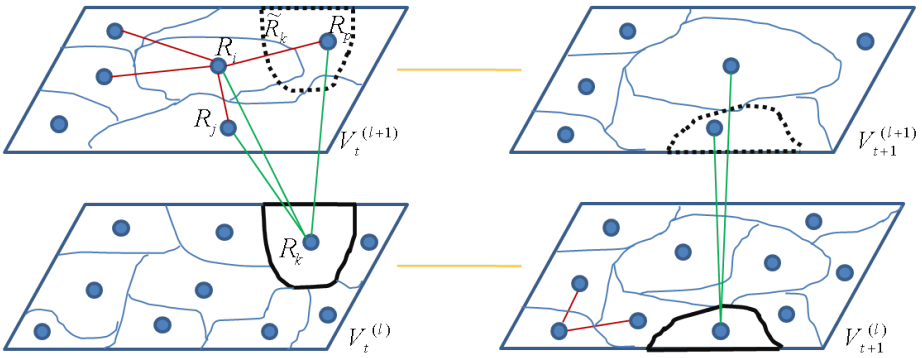
The remainder of this paper is organized as follows. In Section 3, we describe the multi-layer spectral clustering algorithm for video segmentation in detail. Section 3 presents the experimental results. Section 4 concludes the paper.

2 Multi-layer Spectral Clustering for Video Segmentation

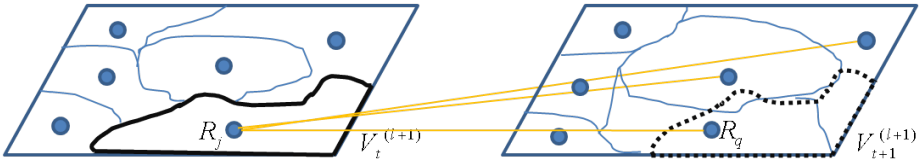
2.1 Multi-layer graph model

For graph-based segmentation algorithms, the final segmentation quality mainly depends on the graph structure and affinities (edge weights). Therefore, to produce high-quality segmentation results with object-level details and frame-to-frame consistence, it is very important to define graph affinities by integrating spatial and temporal grouping cues.

Traditional image segmentation approaches usually define a graph $G = (V, E)$, with V being the pixel set and the weights on E being appearance similarities between pixels. Here, we design a new multi-layer graph $G^* = (V^*, E^*)$, with each layer corresponding to an initial segmentation result under specific granularity. The nodes $V^* = \bigcup\{V_t^{(l)}\}_{t=1,\ldots,T}^{l=1,\ldots,L}$ are the union of multi-parametric presegment sets of a video clip (containing 2 frames as an simple illustration in this paper), where the node subset $V_t^{(l)}$ corresponds to $N_t^{(l)}$ image segments of frame t in layer l. We use mean shift algorithm [16] with varying parameters to get the presegments of all frames [17], which exactly construct the node set V^* in the multi-layer graph.



(a) Connections within frame and between layers.



(b) Connections between consecutive frames.

Fig. 2. The proposed multi-layer graph model. (This figure is better viewed in color.)

In graph G^*, there are three types of undirected edges $\{e_{ij}^* \in E^*\}$, as shown the red, green and orange lines in Fig.2, with the corresponding affinities defined respectively as follows:

1. **Intra-frame neighborhood affinity**. Regions (i.e. presegments as mentioned above) sharing a common boundary within a frame at the same layer are connected (marked by red lines in Fig.2(a)). Let R_i and R_j are intra-frame neighboring regions, and $f_c(R_i)$ and $f_c(R_j)$ are their color histograms respectively. The similarity between intra-frame neighboring regions, R_i and R_j, is computed as a Gaussian function of the χ^2 distance between the two

histogram features, i.e.,

$$w_{ij} = \exp(-\theta \cdot D_\chi(f_c(R_i), f_c(R_j))), R_j \in N^s(R_i).$$

where θ is a constant controlling the strength of the weight, and $N^s(R_i)$ represents the spatial neighborhood of region R_i.

2. **Inter-layer neighborhood affinity**. Regions from the same frame but in adjacent layers may be neighbors as well. Let R_k denote the region with black bold boundary in $V_t^{(l)}$, as shown in Fig.2(a). We shift R_k to the frame in $V_t^{(l+1)}$ and get its spatial correspondence region $\tilde{R}_k$ (Here, ˜ means nonexisting region) in $V_t^{(l+1)}$, which is specified with black dashed boundary, intersecting with three regions, i.e., R_i, R_j, R_p. Then, these three regions are considered as in the inter-layer neighborhood of R_k, denoted as $N^l(R_k)$. Thus, we connect them by green edges. We compute the affinity between a lower-layer region R_k and its neighboring higher-layer region, R_p for example, as the overlapping percentage, i.e.,

$$w_{kp} = \frac{|\tilde{R}_k \cap R_p|}{|R_k|},$$

where $|R_k|$ is the number of pixels in region R_k. This affinity measure the consistency of presegmentations with different parameters.

3. **Inter-frame neighborhood affinity**. Regions at the same layer but in consecutive frames may have temporal overlap, thus can be connected by edges representing the temporal relation (marked by orange lines in Fig.2(b)). Taking the region R_j with black bold boundary in $V_t^{(l+1)}$ (as shown in Fig.2(b)) as an example, we compute the optical flow and estimate its location in the next frame $V_{t+1}^{(l+1)}$ as region with black dashed boundary, which overlaps with three regions. Thus, the three regions are in the temporal neighborhood of region R_j, i.e., $\in N^t(R_j)$, and connected with R_j respectively using orange edges.

   To compute the affinity between neighboring regions in consecutive frames, we resort to similarity between contour shapes [18], as shapes of the neighboring region boundaries convey important information of whether they are from the same object or not. One of the most popular shape descriptors is shape context which is mainly used to recover point correspondences. In this paper, we make use of the distribution over the tangent angles of all boundary points of region R_i as its shape feature, denoted as histogram $f_s(R_i)$. This shape descriptor is proved robust and discriminative enough for our region matching problem. Fig.3 gives some examples. The similarity between R_j and its temporal neighbor, R_q for example, is computed as

$$w_{jq} = \exp(-\gamma \cdot D_\chi(f_s(R_j), f_s(R_q))), R_q \in N^t(R_j).$$

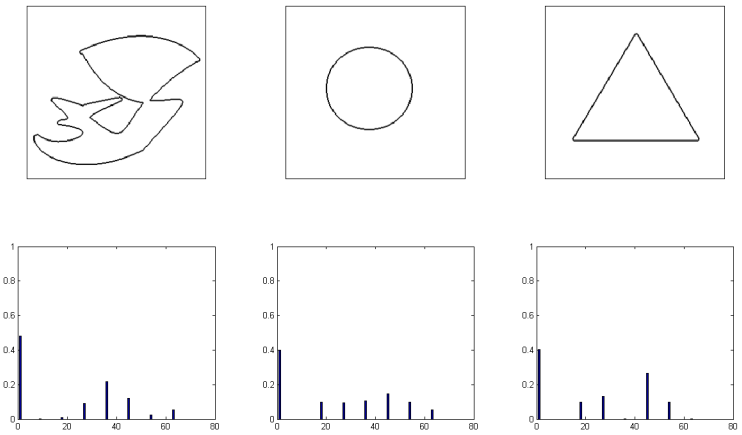


Fig. 3. Discriminative power of the shape descriptor. Some example shapes in the top row and their corresponding shape histograms in the second row.

To summarize, the edge weights $w^*_{ij}, i, j \in V^*$ in the multi-layer graph can be written as

$$w^*_{ij} = \begin{cases} t_1 \exp(-\theta \cdot D_\chi(f_c(R_i), f_c(R_j))) & R_j \in N^s(R_i) \\ t_2 \frac{|\tilde{R}_i \cap R_j|}{|R_i|} & R_j \in N^l(R_i) \\ t_3(1 - \frac{1}{2} D_\chi(f_s(R_i), f_s(R_j))) & R_j \in N^t(R_i) \\ 0 & otherwise \end{cases} \tag{1}$$

Here, t_1, t_2 and t_3 are parameters to balance the three types of weights, with $t_1 + t_2 + t_3 = 1$. Note that the parameters θ and γ in spatial and temporal similarity measures vanish away due to redundancy (with t_1 an t_3 respectively).

2.2 Spectral Segmentation

Once we obtain the affinity matrix for the multi-layer graph $W = [w^*_{ij}]$, its degree matrix D is a diagonal matrix with $d_{ii} = \sum_j W_{ij}$, and the normalized affinity matrix $L = D^{-1}W$. Instead of adopting traditional spectral clustering algorithm, we make use of a more efficient and scalable method, named Power Iteration Clustering (PIC), to get the final segmentation result. To compute the largest eigenvector of a matrix, Power Iteration method starts with an arbitrary vector $v^0 \neq 0$ and repeatedly performs the update $v^{t+1} = \frac{Lv^t}{||Lv^t||_1}$. It has been proved that the intermediate vector v^t, which stops after it has converged within clusters but before final convergence, is an extremely good clustering indicator. We define the velocity at t as $\delta^t = v^t - v^{t-1}$, then the acceleration is $\epsilon^t = \delta^t - \delta^{t-1}$. The power iteration stops when $|\epsilon^t| \cong 0$.

The power iteration clustering method assigns a label $k \in \{1, ..., K\}$, where K is a user-defined parameter, to each presegment in our multi-layer graph. The final segmentation is obtained from the clustering result of the lowest subgraph, which is more coherent and keeps necessary details. The overall procedure is summarized in Algorithm 1.

Algorithm 1 multi-layer spectral segmentation algorithm

Input: Video clip, and the number of clusters K.
Output: Video segmentation result.

1. Construct a multi-layer graph G^*, calculate the weight matrix W^* according to Eq.(1), and compute D and L.
2. Pick an initial vector v^0, repeat set $v^{t+1} = \frac{Lv^t}{||Lv^t||_1}$ and $\delta^{t+1} = |v^{t+1} - v^t|$, increment t, untill $|\delta^t - \delta^{t-1}| \cong 0$.
3. Cluster v^t into K clusters via K-means clustering algorithm.
4. Assign the presegment region R_i to cluster k if and only if element i of the vector v^t is assigned to cluster k.
5. Output the segmentation result in the lowest layer.

2.3 Scaling Up Method

The multi-layer spectral segmentation algorithm is successful in producing coherent segmentations for a variety of videos. However, as it defines a graph based on multi-layer presegments, there is restriction on the spatial and temporal size of the video that it can process. To overcome this bottleneck, we design a scaling up method. We segment video into several spatial-temporal volumes, which are actually very coarse segmentations. Next, we segment each volume using Algorithm 1. Performing multiple passes over the disjoint volumes, we can obtain an identical result with the original algorithm.

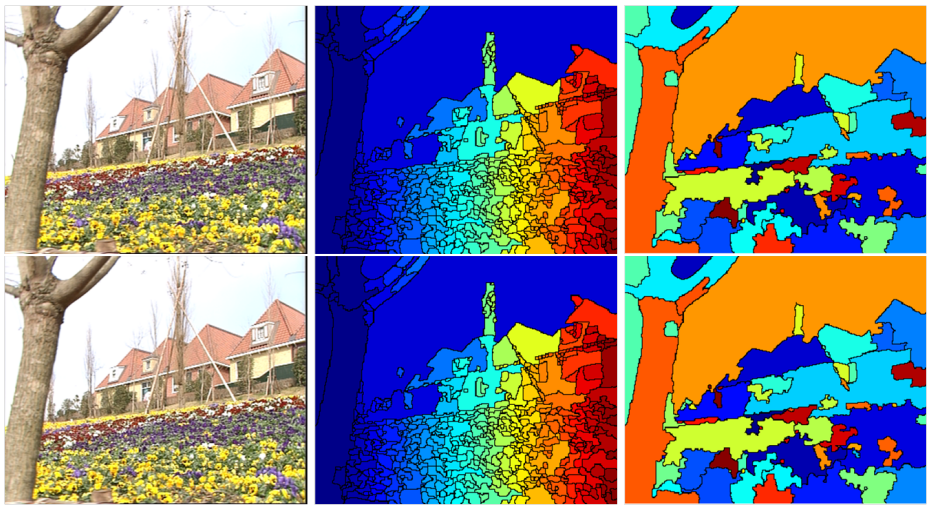
3 Experimental Results

This section presents our quantitative and qualitative evaluation results of the proposed method on a wide range of videos.

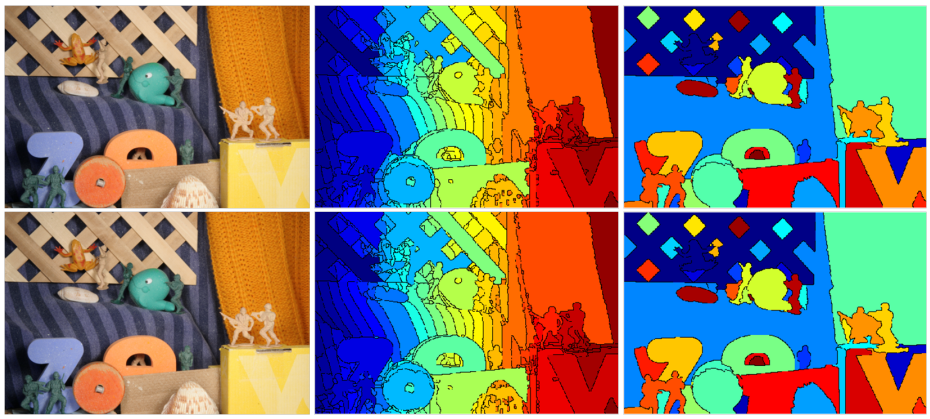
As for parameter setting, the initial mean shift segmentation parameters and the number of total segments K are chosen empirically according to the video contents. t_1, t_2, t_3 can be chosen to put the three kind of affinities into comparable ranges. In our experiments, they are not sensitive within a reasonable variations, so we simply set them all to 1/3 to get the following results.

3.1 Qualitative Results

We compare our algorithm with the baseline segmentation algorithm and the related state-of-the-art methods over a wide range of videos. Some examples are shown below.



(a) Flower Sequence



(b) Army Sequence

Fig. 4. Comparison with [16] on two sequences. From top to bottom rows: original sequence; results of [16]; our segmentation results.

Fig.4 shows the comparison results of our method with the baseline mean shift segmentation algorithm [16]. Apparently, our method gives coherent segmentations on the two videos, while the results of [16] are meaningless. Note that

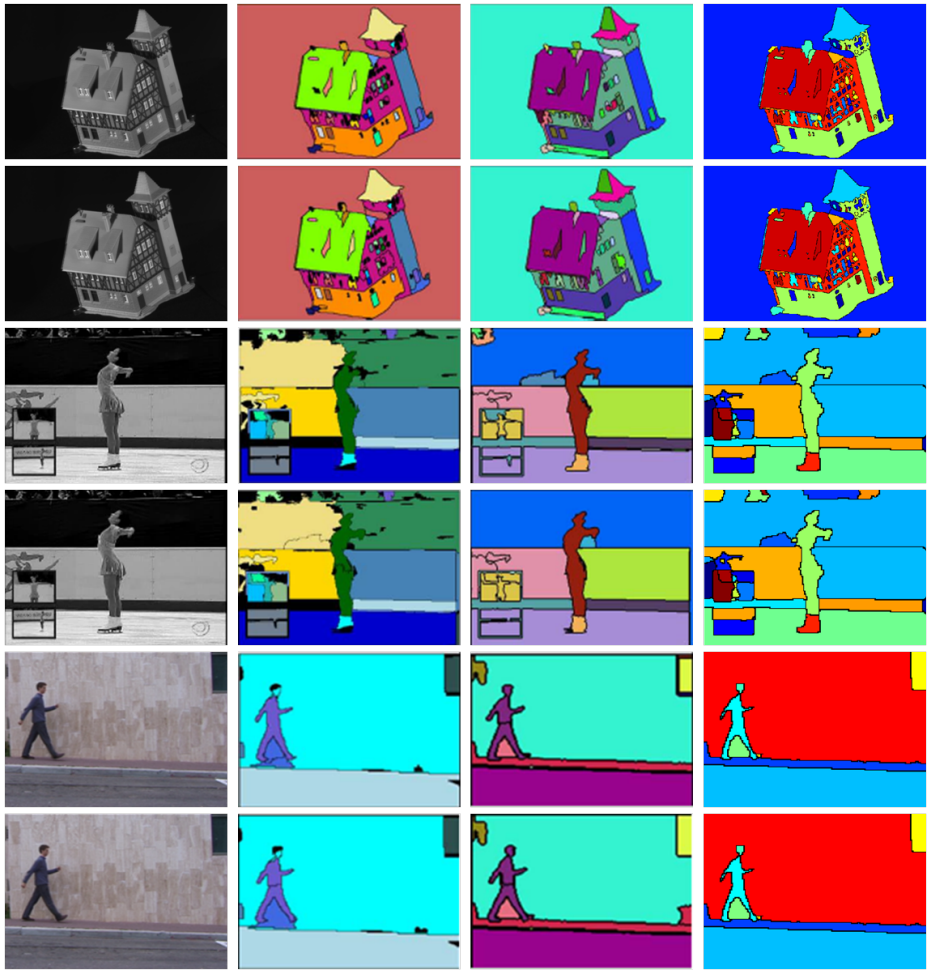


Fig. 5. Comparison with [10] and [8] on House, Kwan, and Walking sequence(listed from top to bottom rows), from left to right columns: original sequence; results of [10]; results of [8]; results of our algorithm.

our method not only identifies regions occupied by different moving objects, but also gives consistent segmentations in the dynamic texture area, e.g. the flower bed in (a).

Fig.5 compares our results against [10] and [8] on three video sequences. Our segmentation retains more important details without losing the overall meaning of the scene: top of the chimney and windows in House, stand in Kwan and the person and background in Walking. Looking at the walking person further, we segment the whole person out, but [10] merges face with background and [8] segment the person into upper and lower body. When it is to the background, our method gives more sound regions of wall and road without outthrust in the results of the other two.

3.2 Quantitative Results

Most existing work on video segmentation presentsonly qualitative evaluation, such as [3] [6] [9] [10]. We can't deny that quantitative evaluation for video segmentation is really a big challenge. The main problem is that there is no unique segmentation of a video clip, and then we could not obtain the convincing ground truth.

Nevertheless, there are some trials on quantitative evaluation on object segmentation in videos [8] [19]. Inspired by these works, we choose *segmentation covering* introduced by [20] to measure the accuracy of object-oriented video segmentation, as did in [19]. Given a ground truth segmentation S and another segmentation S' using some method, the covering of S by S' is defined as

$$C(S' \to S) = \frac{1}{N} \sum_{R \in S} |R| \cdot \max_{R' \in S'} d(R, R'), \tag{2}$$

where N denotes the number of pixels in the image frame. $d(R, R')$ is the Dice coefficient in 3D between the labeled spatio-temporal segments R and R' and its computed as $|R \bigcap R'|/|R \bigcup R'|$.

We use segmentation covering (Equation 2) to quantitatively evaluate our proposed method on human segmentation in Activity videos [21]. This database consists of 10 human activities (including walking, jumping, hand-waving, running, etc.). Each activity includes 9 videos, and each video consists of 30 to 120 frames. In all videos, one person moves in front of a static background. The ground truth of each video segmentation is obtained by manually labeling the moving person and the background. Some sample frames from running sequence and their ground truth are shown in Fig. 6.

In Table 1, we report the comparison results of our method with the baseline Mean-shift segmentation algorithm [16] and the most related method, Video Object Segmentation by Tracking Regions (VOSTR) [8] over 6 activity videos (i.e. jump in place, walk, jump, gallop, wave1, and run). For each video, we give the average segmentation covering and the corresponding standard deviation. As can be seen, our method gives the best average segmentation coverings over all videos with a little sacrifice in the standard deviation. So, considering the two factors together, our approach slightly outperforms the other two.

Fig. 6. Example frames from Activity video running in the top row and their ground truth in the second row.

Table 1. Segmentation covering of the person and its corresponding standard deviation (in parentheses) over frames in some activity videos obtained with mean-shift [16], VOSTR [8], and Multi-layer (ours). The higher segmentation covering and the lower standard deviation, the better. The best segmentation covering of each video is shown in bold.

Method	Mean-shift	VISOR	Multi-layer
eli-pjump	0.64 (±0.07)	0.75 (±0.07)	**0.82** (±0.10)
eli-walk	0.46 (±0.16)	0.35 (±0.23)	**0.50** (±0.26)
ido-jump	0.53 (±0.16)	0.44 (±0.28)	**0.57** (±0.18)
ido-gallop	0.67 (±0.02)	0.49 (±0.25)	**0.78** (±0.07)
lena-wave1	0.85 (±0.03)	0.54 (±0.38)	**0.91** (±0.04)
lyova-run	0.32 (±0.25)	0.07 (±0.06)	**0.44** (±0.23)

4 Conclusion

In this paper, we propose an novel multi-layer spectral clustering algorithm for video segmentation. The advantage of multi-layer graph model lies in providing a non-parametric way to model relationship between presegments. We make use of efficient power iteration clustering to get the final segmentation result and design a simple scale-up method to make the algorithm more applicable for real world problems. Experimental results show the priority of our method to some related state-of-the-art methods.

Acknowledgement. This work is partially supported by National Basic Research Program of China (973 Program) under contract 2009CB320902; and Natural Science Foundation of China (NSFC) under contract Nos. 61025010, 60833013, and 60832004.

References

1. Tron, R., Vidal, R.: A benchmark for the comparison of 3-d motion segmentation algorithms. In: CVPR. (2007)
2. Rao, S., Tron, R., Vidal, R.: Motion segmentation via robust subspace seperation in the presence of outlying, incomplete, or corrupted trajectories. PAMI (2010)
3. DeMenthon, D., Megret, R.: Spatio-temporal segmentation of video by hierarchical mean shift analysis. In: CVPR. (2000)
4. Khan, S., Shah, M.: Object based segmentation of video using color, motion and spatial information. PAMI (2005)
5. Torsello, A., Pavan, M., Pelillo, M.: Object based segmentation of video using color, motion and spatial information. In: EMMCVPR. (2005)
6. Grundmann, M., Kwatra, V., Han, M., Essa, I.: Efficient hierarchical graph-based video segmentation. In: CVPR. (2010)
7. Nagahashi, T., Fujiyoshi, H., Kanade, T.: Video segmentation using iterated graph cuts based on spatio-temporal volumes. In: ACCV. (2009)
8. Brendel, W., Todorovic, S.: Video object segmentation by tracking regions. In: ICCV. (2009)
9. Ke, Q., Kanade, T.: A subspace approach to layer extraction. In: CVPR. (2001)
10. Hedau, V., Arora, H., Ahuja, N.: Matching images under unstable segmentations. In: CVPR. (2008)
11. Kim, T.H., Lee, K.M., Lee, S.U.: Learning full pairwise affinities for spectral segmentation. In: CVPR. (2010)
12. Weiss, Y.: Segmentation using eigenvectors: a unifying view. In: ICCV. (1999)
13. Shi, J., Malik, J.: Normalized cuts and image segmentation. PAMI (2000)
14. Cour, T., Bébézit, F., Shi, J.: Segmentation using eigenvectors: a unifying view. In: CVPR. (2005)
15. Lin, F., Cohen, W.W.: Power iteration clustering. In: ICML. (2010)
16. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. PAMI (2002)
17. Kohli, P., Ladicky, L., P.Torr: Robust higher order potentials for enforcing label consistency. In: CVPR. (2008)
18. Hu, G., Gao, Q.: A non-parametric statistics based method for generic curve partition and classification. In: ICIP. (2010)
19. Vazquez-Reina, A., Avidan, S., Pfister, H., Miller, E.: Multiple hypothesis video segmentation from superpixel flows. In: ECCV. (2010)
20. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: From contours to regions:an empirical evaluation. In: CVPR. (2009)
21. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Action as space-time shapes. IEEE TPAMI (2007)