09-05

15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Event Based Surveillance Video Synopsis Using Trajectory Kinematics Descriptors

Wei-Cheng Wang, Pau-Choo Chung
Dept. of Electrical Engineering
National Cheng Kung University
Tainan, Taiwan
pcchung@ee.ncku.edu

Chun-Rong Huang, Wei-Yun Huang
Dept. of Computer Science and Engineering
National Chung Hsing University
Taichung, Taiwan
crhuang@nchu.edu.tw

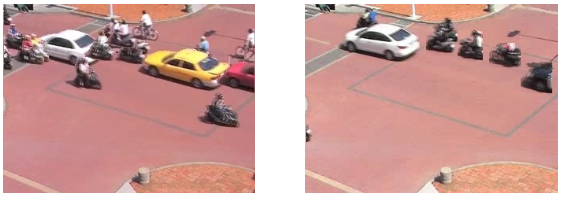
Abstract

Video synopsis has been shown its promising performance in visual surveillance, but the rearranged foreground objects may disorderly occlude to each other which makes end users hard to identify the targets. In this paper, a novel event based video synopsis method is proposed by using the clustering results of trajectories of foreground objects. To represent the kinematic events of each trajectory, trajectory kinematics descriptors are applied. Then, affinity propagation is used to cluster trajectories with similar kinematic events. Finally, each kinematic event group is used to generate an event based synopsis video. As shown in the experiments, the generated event based synopsis videos can effectively and efficiently reduce the lengths of the surveillance videos and are much clear for browsing compared to the states-of-the-art video synopsis methods.

1. Introduction

With the increasing number of surveillance cameras, browsing videos of these cameras for event retrieval costs a lot of human resources. Providing a new browsing way to help end users to effectively and efficiently review the videos in a short time becomes an important issue in visual surveillance. One of the most naïve ways to browse surveillance videos is the time lapse videos constructed by using uniform sampling. However, the moving velocities and behaviors of objects in surveillance videos are different. As a result, motion information of time lapse videos will disappear.

To solve aforementioned problems, video synopsis [1][2][3][4] suggests an alternative approach by rearranging all foreground objects into a condensed video. Although video synopsis can effectively shorten the video lengths for browsing, each frame of the synopsis video will contain a lot of foreground objects with different motion behaviors as shown in Figure 1(a). These crowded foreground objects will lead to occlusion and visibility problems during browsing. To avoid these problems, event based video synopsis is proposed which can reveal foreground objects of the same motion behavior as shown in Figure 1(b). Thus, the event based video synopsis can achieve better visual quality for browsing.



(a) (b)

Figure 1. The sampled frames of (a) the synopsis video and (b) the event based synopsis video.

Recently, Pritch et al. [5] cluster foreground objects which have similar appearances and motions to several groups by using spectral clustering [6]. Based on the clustering results, synopsis videos for each cluster are generated by using Markov random field optimization [7]. Nevertheless, the number of clusters needs to be assigned. Moreover, the appearance and motion features are easily affected by the lengths of trajectories of foreground objects. Their method also suffers from the same computational complexity problem as [1]. To obtain the number of groups in advance, Chou et al. [8] consider the numbers of entrances and exits to obtain event groups. The distances between trajectories are defined by the longest common subsequence [9]. Hierarchical clustering is adopted to group trajectories of similar events for video synopsis. Nevertheless, the entrances and exits are not generally available for outdoor environments.

In this paper, a novel event based surveillance video synopsis method is proposed which includes trajectory kinematics representation, trajectory event clustering, and synopsis video generation. To effectively represent the kinematic properties of trajectories, trajectory kinematics descriptor (TKD) [10] is applied. With the similarities between TKDs of trajectories, affinity propagation [11] is then applied to cluster trajectories of similar kinematic events. To achieve real-time performance, a synopsis table introduced in [4] is applied which can efficiently rearrange foreground objects with similar kinematic events to generate the event based synopsis videos. As shown in the experiments, our method can successfully generate the event based synopsis videos in real-time with better visual readability.

2. Method

2.1. Trajectory Representation

To represent each trajectory and compare the distance between two trajectories, trajectory kinematics descriptor (TKD) [10] is applied, which can overcome the length problem of trajectory comparison. TKD assumes that each foreground object moves in a three-dimensional

Euclidean space $\Re^3$ and imposes Frenet-Serret frames [11] to construct the descriptor of the trajectory.

After foreground object extraction and tracking [4], a trajectory $\boldsymbol{r}_i$ of the ith foreground object O_i is represented as $\boldsymbol{r}_i = \{\boldsymbol{r}_i(t_s),\ldots, \boldsymbol{r}_i(t),\ldots, \boldsymbol{r}_i(t_e)\}$ in the Euclidean space, where $\boldsymbol{r}_i(t)$ represents the image center positions of O_i in frame t, and t_s and t_e are the first and latest appearing frames of $\boldsymbol{r}_i$. Please note that t_s and t_e of each trajectory can be different. Let $s_i(t)$ be the arc length of $\boldsymbol{r}_i$ defined as:

$$s_i(t) = \int_{t_s}^{t} \|\boldsymbol{r}_i'(\sigma)\| d\sigma, \tag{1}$$

where $\boldsymbol{r}_i'(\cdot)$ is the velocity vector. TKD applies the Frenet-Serret frame from the point of view of the foreground object to describe the kinematic properties. The tth tangent unit vector $\boldsymbol{T}(t)$, normal unit vector $\boldsymbol{N}(t)$, and binormal unit vector $\boldsymbol{B}(t)$ unit vectors of the tth Frenet-Serret frame are defined as follows:

$$\boldsymbol{T}_i(t) = \frac{d\boldsymbol{r}_i(t)}{ds_i(t)}, \tag{2}$$

$$\boldsymbol{N}_i(t) = \frac{\frac{d\boldsymbol{T}_i(t)}{ds_i(t)}}{\left\|\frac{d\boldsymbol{T}_i(t)}{ds_i(t)}\right\|}, \tag{3}$$

and

$$\boldsymbol{B}_i(t) = \boldsymbol{T}_i(t) \times \boldsymbol{N}_i(t). \tag{4}$$

The unit vectors $\boldsymbol{T}_i(t)$, $\boldsymbol{N}_i(t)$, and $\boldsymbol{B}_i(t)$ form an orthonormal basis of $\Re^3$. Then, the normalized histograms $\boldsymbol{h}(\boldsymbol{T}_i)$, $\boldsymbol{h}(\boldsymbol{N}_i)$ and $\boldsymbol{h}(\boldsymbol{B}_i)$ of the three unit vectors $\boldsymbol{T}_i$, $\boldsymbol{N}_i$ and $\boldsymbol{B}_i$ are constructed. The TKD $\boldsymbol{h}(\boldsymbol{r}_i)$ of $\boldsymbol{r}_i$ is defined by composing $\boldsymbol{h}(\boldsymbol{T}_i)$, $\boldsymbol{h}(\boldsymbol{N}_i)$ and $\boldsymbol{h}(\boldsymbol{B}_i)$ as follows:

$$\boldsymbol{h}(\boldsymbol{r}_i) = [\boldsymbol{h}(\boldsymbol{T}_i)\ \boldsymbol{h}(\boldsymbol{N}_i)\ \boldsymbol{h}(\boldsymbol{B}_i)]. \tag{5}$$

With TKD, each trajectory $\boldsymbol{r}_i$ is transferred to a descriptor $\boldsymbol{h}(\boldsymbol{r}_i)$ which has the same length for comparison.

2.2. Event Clustering

Because TKD represents the kinematic properties of trajectories, they can be used to identify trajectories of similar kinematic events. To group trajectories of similar kinematic events, we apply affinity propagation [11] with TKDs. Affinity propagation considers each TKD as an exemplar. To find clusters, two kinds of messages, responsibility and availability, sent from each TKD to the remaining TKDs are used. By updating the messages, the current affinity reflects the score that one TKD prefers another TKD as its exemplar. Unlike most clustering algorithms, a fixed number of potential clusters is not required in advance. To perform clustering, the affinity matrix $\boldsymbol{s}$ is constructed at first. Each element $s(i, j)$ of $\boldsymbol{s}$ represents the similarity between two trajectories $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$ and is defined as follows:

$$s(i,j) = e^{-\|\boldsymbol{h}(\boldsymbol{r}_i)-\boldsymbol{h}(\boldsymbol{r}_j)\|} + e^{-d(\boldsymbol{r}_i,\boldsymbol{r}_j)}, \tag{6}$$

where $\|\boldsymbol{h}(\boldsymbol{r}_i) - \boldsymbol{h}(\boldsymbol{r}_j)\|$ is the distance between two TKDs, and $d(\boldsymbol{r}_i, \boldsymbol{r}_j)$ is the difference between the positions of the first and latest appearing frames of $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$, respectively. Here, $d(\boldsymbol{r}_i, \boldsymbol{r}_j)$ is defined as follows:

$$d(\boldsymbol{r}_i,\boldsymbol{r}_j) = \|r_i(t_{si}) - r_j(t_{sj})\| + \|r_i(t_{ei}) - r_j(t_{ej})\|. \tag{7}$$

Based on $s(i, j)$, the responsibility $r(i, j)$ between trajectories $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$ is defined as follows:

$$r(i,j) = s(i,j) - \max_{j' s.t. j' \neq j}\{a(i,j') + s(i,j')\}. \tag{8}$$

If two trajectories are similar, $r(i, j)$ will become larger which indicates that $\boldsymbol{r}_i$ is suitable to represent $\boldsymbol{r}_j$ based on the similarity. The availability $a(i, j)$ of $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$ is defined as follows:

$$a(i,j) = \min\left\{0, r(j,j) + \sum_{i' s.t. m' \notin \{i,j\}} \max\{0, r(i',j)\}\right\}. \tag{9}$$

Finally, the self-availability is defined as:

$$a(j,j) = \sum_{i' s.t. i' \neq j} \max\{0, r(i',j)\}. \tag{10}$$

By continuously exchanging the responsibility and availability messages between TKDs of different foreground objects, the corresponding exemplar of $\boldsymbol{r}_j$ is obtained by

$$j^* = \underset{j \neq i}{\operatorname{argmax}}\{a(i,j) + r(i,j)\}, \tag{11}$$

where $\boldsymbol{r}_{j^*}$ and $\boldsymbol{r}_i$ belong to the same event group. As a result, events of trajectories with similar TKDs will be clustered to the same group.

2.3. Event based Surveillance Video Synopsis

After affinity propagation, the trajectories in the video are separated into K groups and represented as $G = \{G_1, \ldots, G_k, \ldots, G_K\}$, where G is the union of all groups and G_k is the kth group of G. Because each group G_k contains trajectories of similar kinematic events, we generate the synopsis video SV_k for each G_k. To effectively arrange the temporal positions of the foreground objects in G_k, we construct a synopsis table T_k [4], which stores the latest occupied time slots for each pixel (x, y) in the synopsis video.

Because a trajectory represents continuous spatial and temporal changes of a foreground object, the appearing order of the instances of the trajectory is fixed to avoid the discontinuity or jittering of the foreground object in the synopsis video. Two situations need to be considered during arranging $\boldsymbol{r}_i$ to SV_k including that $\boldsymbol{r}_i$ is firstly arranged in SV_k and a part of $\boldsymbol{r}_i$ is already arranged in SV_k, respectively. The synopsis table T_k here is used to record the aforementioned information. Let the initial values of elements of T_k be 0. The temporal location $TL(\boldsymbol{r}_i(t))$ of $\boldsymbol{r}_i(t)$ is defined as follows:

$$TL(\boldsymbol{r}_i(t)) = \begin{cases} s', & \text{if } \boldsymbol{r}_i \text{ is firstly arranged} \\ TL(\boldsymbol{r}_i(t-1)) + 1, & \text{otherwise} \end{cases}, \tag{12}$$

where s' is the latest available temporal location, which will not occlude previous appearing foreground objects in SV_k and is defined as follows:

$$s' = \max_{x,y}\ \{T_k(x,y) \mid \forall (x,y) \in F(\boldsymbol{r}_i(t))\} + 1, \tag{13}$$

where $F(\boldsymbol{r}_i(t))$ represents the union of pixels of the foreground object in the trajectory $\boldsymbol{r}_i(t)$, and (x, y) is the pixel location. Because s' is located after the latest occupied location, $\boldsymbol{r}_i$ will not occlude previously appearing foreground objects. If $\boldsymbol{r}_i(t-1)$ is arranged in SV_k, $\boldsymbol{r}_i(t)$ needs to appear at the next frame of $\boldsymbol{r}_i(t-1)$ to avoid the foreground fragments. After assigning the temporal location for each element of the trajectory, T_k is updated to record the most recent occupied time slots for (x, y) in SV_k as follows:

$$T_k(x,y) = \begin{cases} TL(\boldsymbol{r}_i(t)), & \forall (x,y) \in F(\boldsymbol{r}_i(t)) \\ T_k(x,y), & \text{otherwise} \end{cases}. \tag{14}$$

Based on the synopsis table, the temporal locations of each trajectory can be computed without nonlinear optimization as shown in [1][2]. Then, each foreground object is sequentially stitched on the synopsis video based on the temporal location of Eq. (14) and its trajectory. Please note that we generate a synopsis video SV_k for each group G_k to provide users browsing the synopsis videos with different kinematic events.

3. Experimental Results

3.1. Datasets

We used four surveillance videos in [4] for evaluation including three outdoor and one indoor videos. The resolutions of these videos are 320×240 and the numbers of frames are shown in Table 1. For quantitative comparisons, the frame reduction rate (FR) is defined as:

$$FR = \frac{\sum_{k=1}^{K} \#F(SV_k)}{\#F(V)}, \tag{15}$$

where $\#F(SV_k)$ and $\#F(V)$ are the numbers of frames of SV_k and V, respectively, and K is the number of clusters obtained by affinity propagation [11]. The method is implemented on an Intel i7 3.4-GHz CPU with 16-GB memory computer.

3.2. Results

Table 1 shows the numbers of objects and frames of four sampled event based synopsis videos. The FR and average frame per second (FPS) for each evaluation video are shown in Table 2. Because each event based synopsis video contains few foreground objects with the similar kinematic behavior compared to [4], the FR of the proposed method is less than that of [4]. The FPS of the proposed method is lower than that of [4], because of the computation of TKDs and affinity propagation.

Figure 2(a) shows the synopsis results of [4]. Different kinds of kinematic events, i.e. vehicles move to different directions, disorderly appear in the video frame. In contrast, the proposed method can generate an event based synopsis video for the higher velocity vehicle flow moving from the left side to the right side based on the clustering results as shown in Figure 2(b). Figure 2(c) shows the lower velocity vehicle flow but the same moving direction as the vehicles in Figure 2(b). Figure 2(d) shows the vehicle flow from the right side to the left side with similar velocities. Finally, Figure 2(e) shows the turning vehicle flow. Thus, different kinds of kinematic events can be revealed in different synopsis videos for browsing compared to traditional video synopsis methods. Figure 3(a) show all of the different kinds of kinematic events in the same frame of [4]. In Figure 3(b), the pedestrians walk toward to the upper right of the scene in the event based synopsis video. The pedestrians in Figure 3(c) walk from the upper right of the scene to the bottom of the scene. Figure 3(d) and (e) show the different moving directions of pedestrians. These cases show the effectiveness of TKD in distinguishing behaviors. Figure 4(b) and (c) show individual pedestrians and crowd pedestrians. Figure 4(d) shows the pedestrians moving in the opposite direction compared to Figure 4(b) and (c). Because motorcycles and bicycles have faster moving velocities compared to pedestrians, they belong to a different kinematic event group and are shown in a different event based synopsis video in Figure 4(e). In Figure 5, different kinds of moving directions of pedestrians are revealed in different event based synopsis videos. Due to limited space, please refer to the demo video, which is available at http://cvml.cs.nchu.edu.tw/EventVideoSynopsis.htm.

Table 1. The number of objects and frames of the event based synopsis videos.

Video	Original	Synopsis Video	
	# of Frames	# of Objects	# of Frames
Crossroad	70195	48	425
		55	305
		43	250
		23	469
Street	79449	210	2863
		159	3815
		118	1742
		227	9313
Sidewalk	104864	47	1414
		76	5328
		95	520
		109	1543
Hall	66771	61	2074
		44	2617
		53	976
		39	470

Table 2. The comparison metrics.

Video	[4]		Proposed	
	FR	FPS	FR	FPS
Crossroad	0.181	71.43	0.021	52.73
Street	0.237	66.67	0.236	84.61
Sidewalk	0.215	71.43	0.058	18.60
Hall	0.174	71.43	0.132	26.04

4. Conclusions

In this paper, we propose a new event based video synopsis method, which can separate foreground objects with different kinematic events to different synopsis videos. The results can reduce the burden of end users when

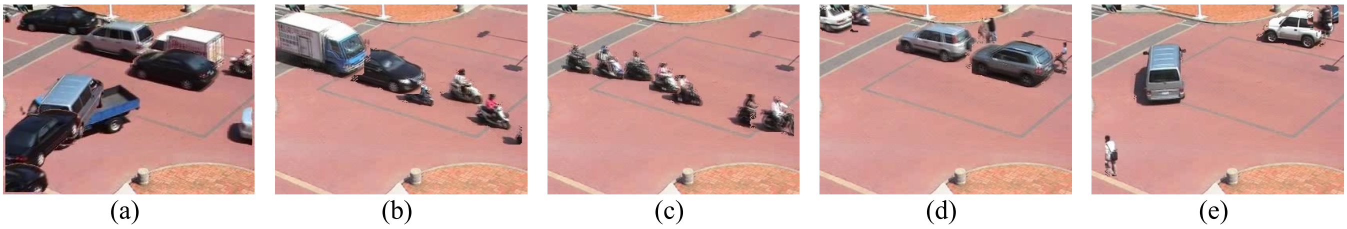


Figure 2. Results of the crossroad video. (a) Synopsis video in [4], (b) SV_1 of G_1, (c) SV_2 of G_2 , (d) SV_3 of G_3 , (e) SV_4 of G_4.

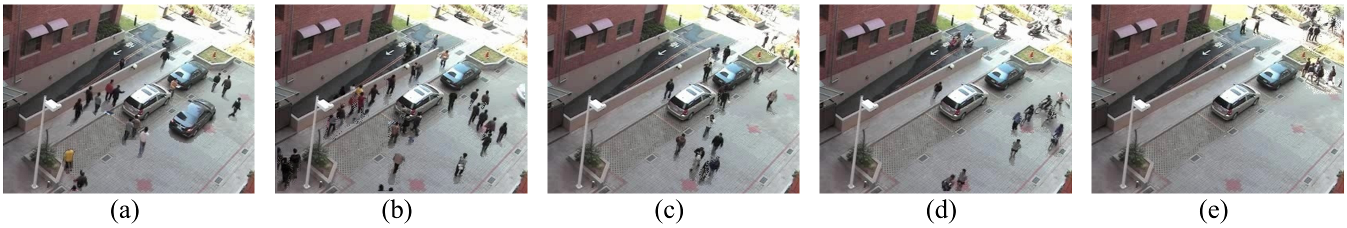


Figure 3. Results of the street video. (a) Synopsis video in [4], (b) SV_1 of G_1, (c) SV_2 of G_2 , (d) SV_3 of G_3 , (e) SV_4 of G_4.

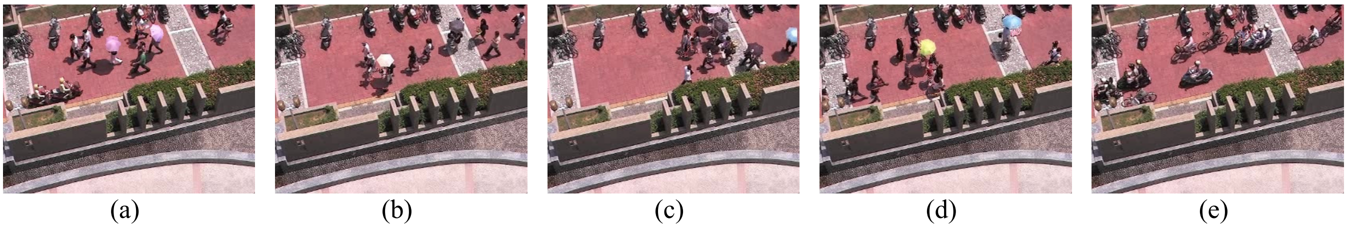


Figure 4. Results of the sidewalk video. (a) Synopsis video in [4], (b) SV_1 of G_1, (c) SV_2 of G_2 , (d) SV_3 of G_3 , (e) SV_4 of G_4.

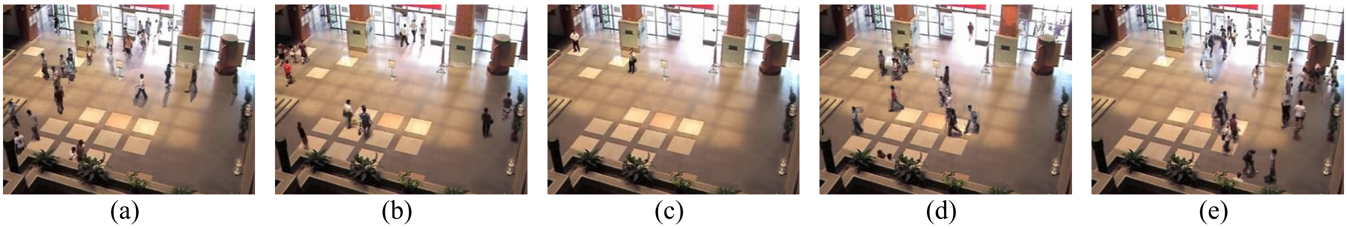


Figure 5. Results of the hall video. (a) Synopsis video in [4], (b) SV_1 of G_1, (c) SV_2 of G_2 , (d) SV_3 of G_3 , (e) SV_4 of G_4.

browsing the synopsis videos and make the users easily focus on specific events. In the future, we will apply more high level descriptors of foreground objects and scenes to provide event based synopsis videos which comply with human semantics.

Acknowledgments

This work was supported by the Ministry of Science and Technology of Taiwan under Grant 104-2221-E-005-027-MY3 and Grant MOST103-2923-E-006-001-MY3.

References

[1] Y. Pritch, A. Rav-Acha, and S. Peleg, "Nonchronological Video Synopsis and Indexing", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1971–1984, 2008.

[2] S. Feng, Z. Lei, D. Yi, and S. Z. Li, "Online Content-Aware Video Condensation," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2082–2087, 2012.

[3] X. Li, Z. Wang and X. Lu, "Surveillance Video Synopsis via Scaling Down Objects," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 740–755, Feb. 2016.

[4] C.-R. Huang, P.-C. Chung, D.-K. Yang, H.-C. Chen, and G.-J. Huang, “Maximum a Posteriori Probability Estimation for Online Surveillance Video Synopsis,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 8, pp. 1417–1429, 2014.

[5] Y. Pritch, S. Ratovitch, A. Hendel, and S. Peleg, "Clustered Synopsis of Surveillance Video", in *Proc. IEEE Intl. Conf. on Advanced Video and Signal Based Surveillance*, pp. 195–200, 2009.

[6] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, 2000.

[7] V. Kolmogorov and R. Zabih, “What Energy Functions Can be Minimized via Graph Cuts?” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 147–159, Feb. 2004.

[8] C-L. Chou, C.-H. Lin, T.-H. Chiang, H.-T. Chen, and S.-Y. Lee, "Coherent Event-based Surveillance Video Synopsis Using Trajectory Clustering," in *Proc. Intl. Conf. on Multimedia & Expo Workshops*, pp.1–6, 2015.

[9] B. T. Morris, and M. M. Trivedi, “A Survey of Vision-based Trajectory Learning and Analysis for Surveillance,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no, 8, pp. 1114–1127, 2008.

[10] W.-C. Wang, P.-C. Chung, H.-W. Cheng, and C.-R. Huang, “Trajectory Kinematics Descriptor for Trajectory Clustering in Surveillance Videos,” in *Proc. Intl. Symp. on Circuits and Systems*, pp.1198–1201, 2015.

[11] B. J. Frey and D. Dueck, "Clustering by Passing Messages between Data Points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[12] C. Qu, “Invariant Geometric Motions of Space Curves,” *Lecture Notes in Computer Science: Computer Algebra and Geometric Algebra with Applications*, vol. 3519, pp 139–151, 2005.