Virtual Gazing in Video Surveillance

Yingzhen Yang
Carnegie Mellon University
4720 Forbes Ave.
Pittsburgh, PA 15213, USA
yingzhe1@andrew.cmu.edu

Yang Cai
Carnegie Mellon University
4720 Forbes Ave.
Pittsburgh, PA 15213, USA
ycai@cmu.edu

ABSTRACT

Although a computer can track thousands of moving objects simultaneously, it often fails to understand the priority and the meaning of the dynamics. Human vision, on the other hand, can easily track multiple objects with saccadic motion. The single thread eye movement allows people to shift attention from one object to another, enabling visual intelligence from complex scenes. In this paper, we present a motion-context attention shift (MCAS) model to simulate attention shifts among multiple moving objects in surveillance videos. The MCAS model includes two modules: The robust motion detector module and the motion-saliency module. Experimental results show that the MCAS model successfully simulates the attention shift in tracking multiple objects in surveillance videos.

Categories and Subject Descriptors

H.1.2 [**MODELS AND PRINCIPLES**]: User/Machine Systems –Human information processing.

General Terms

Algorithms, Experimentation.

Keywords

Virtual Gazing, Motion-Context Attention Shift, Motion Detector, Simulation, Motion-Saliency Module.

1. INTRODUCTION

Visual attention is a rare resource in human information processing. Humans instinctively manage attention for complex dynamic scenes with ease. Psychologists have studied the human gaze for the past half century [17]. It is shown that eye movement reveals user behaviors in different tasks such as browsing the web or watching television [1, 2, 3, 8]. Moreover, the gaze fixation in between the movements can be used to trigger actions [4, 5] or infer the user task [6], and multimodal user interfaces [9]. The gaze tracking data is also analyzed to predict the skill levels between different users in [7], and the spatial selective attention is used for visual tracking in [20].

Most gaze studies are based in still images [18]. Recently, gaze in video information processing is growing interest in many disciplines. For example, the awareness test [19] shows that people track multiple moving objects in a video with limited attention.

Recent work includes gaze attention behaviors for crowds [10], where a set of gaze constraints based on parameters such as distance or speed to all other characters or objects in the scene determine where and when each character should look. Moreover, an eye gaze model is presented in [11] for an embodied conversational agent, and it can generate different gaze behaviors to stimulate several personalized gaze habits of an embodied conversational agent.

However, how to simulate human gaze in a realistic dynamic scene with multiple objects is still a challenge. Our objective is to model and simulate human gaze shifts with multiple moving objects. We call the simulation of human gaze shift is 'virtual gazing'.

Since we only concentrate on simulating attention shifts among moving objects, the first step of the MCAS model is to find the tracks of moving objects by the robust foreground object detector in [12] and mean-shift based object tracking algorithm in [13]. In the second step, we use a motion-saliency module based on the surprise model [14, 15, 16] to calculate the motion-saliency value for all the locations in the object trajectories and select the location with the highest motion-saliency value as the simulated gaze positions. These simulated gaze positions are called motion-critical positions in the flowing text. In following sessions, we will illustrate the MACS model in more details.

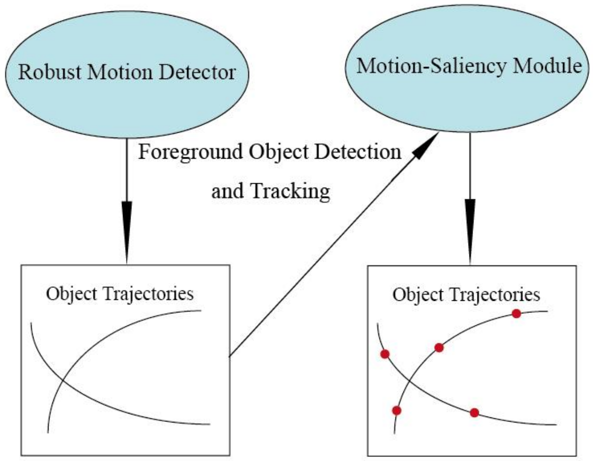

Figure 1. Illustration of the two modules in the MCAS model. The red points in the object trajectories indicate the motion critical locations.

SMVC'10, October 29, 2010, Firenze, Italy.

2. MOTION-CONTEXT ATTENTION SHIFT MODEL

MCAS is comprised of two modules: the robust motion detector and the motion-saliency module. Figure 1 shows the two modules working together to localize the motion-critical positions.

2.1 Robust Motion Detector

The robust motion detector is to find the objects with salient movement, and it obtains the trajectories of moving objects in two successive stages. In the first stage, we detect and segment foreground objects from a video which contains a stationary background. Our approach is based on Bayesian decision model [12] for classification of background and foreground from selected feature vectors is formulated as below:

$$P(C \mid V_t, s) = \frac{P(V_t \mid C, s) P(C \mid s)}{P(V_t \mid s)} \qquad (1)$$

where V_t denotes the feature vector extracted from an image sequence at the pixel s and time instant t, and $C \in \{b, f\}$ is a two-value variable indicating background or foreground. Also, we have:

$$P(V_t \mid s) = P(V_t \mid b, s) P(b \mid s) + P(V_t \mid f, s) P(f \mid s) \qquad (2)$$

According to Bayes decision rule, a pixel is classified as background if the feature vector satisfies

$$P(b \mid V_t, s) > P(f \mid V_t, s) \qquad (3)$$

From (1), (2), (3) becomes:

$$2P(V_t \mid b, s) P(b \mid s) > P(V_t \mid s) \qquad (4)$$

So that if we learn the prior probability $P(b \mid s)$, $P(V_t \mid s)$ and $P(V_t \mid b, s)$ beforehand, we are capable of classifying a feature V_t as either foreground or background.

In the second stage, we use an object tracking method based on mean-shift to obtain the motion trajectories of moving foreground objects by the results of the first stage. This object tracking method iteratively finds the most probable position of a target in the next frame based on its current position, by minimizing the coefficient through mean-shift iterations [13]. By virtue of this robust motion detector module, we can have the motion trajectories of the foreground objects in a video with a stationary background. It is important for the simulation because the motion-critical positions will be found in these trajectories.

2.2 Motion-Saliency Module

2.2.1 *Motivations*

We hypothesize that human attention is correlated with motion saliency, in many cases, surprises in shape, size and speed. For example, a big, unknown, and fast moving object would instinctively attract more attention over small, familiar and slow moving objects. The reflective act is sub-conscious without semantic understanding. We can anticipate the reactive attention shifts based on the motion saliency (shape, size and speed). However, we may not do a long-term prediction of the attention shifts merely based historical data. Instead, our model is just for very short-term, within a half second, depending on the motion context of multiple moving objects.

2.2.2 *Formulation of Motion-Saliency Module*

First, we define a motion map M_p which records the motion history of all the foreground moving objects as below:

$$M_p(t) = [Pos_1, Pos_2, ...Pos_k] \qquad (5)$$

where t is the time instant, k is the number of moving objects at time t, and $Pos_i, i \in \{1, 2...k\}$ denotes the two dimensional coordinates of the i-th moving object on the screen, which is determined by the robust motion detector module. So $M_p(t)$ denotes the position of all moving objects at time t.

Next, under the Bayesian probability framework [14, 15], we give the probabilistic formulation of the motion-saliency model. Suppose $MH(t)$ is the motion history of all moving objects before time t, that is, $MH(t) = \{M_p(k), k = 1, 2, ...t-1\}$.

According to the Bayesian theorem, for each component of the motion record $M_p(t)$, i.e. Pos_i, $i \in \{1, 2...k\}$ we have:

$$P(MH(t) \mid Pos_i) = \frac{P(Pos_i \mid MH(t))}{P(Pos_i)} P(MH(t)) \qquad (6)$$

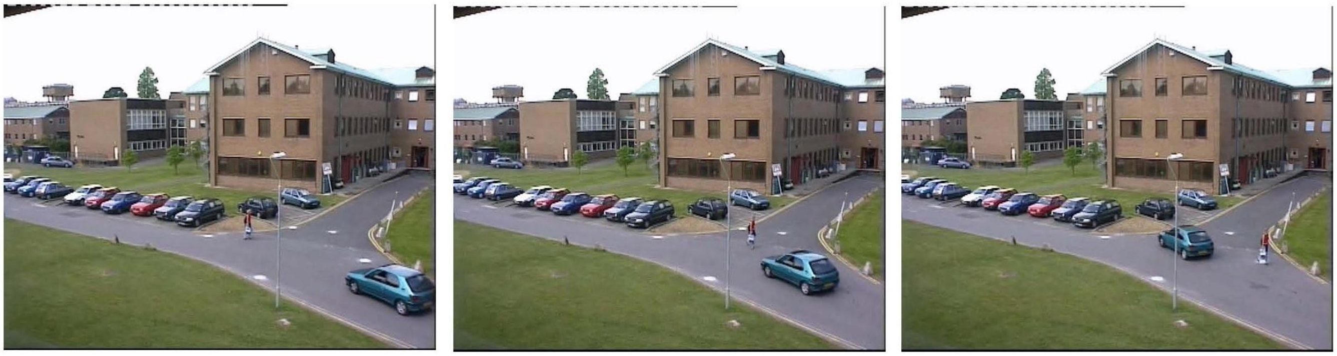

Figure 2. Snapshots of the video clip built from PETS 2001.

Here, $P(MH(t))$ is the prior distribution before time t, and $P(MH(t) \mid Pos_i)$ is the posterior distribution.

If $P(MH(t) \mid Pos_i)$ has little or no difference from $P(MH(t))$, then the movement of the i-th object on screen at time t is not surprising with respect to the motion history of all the moving objects before time t; otherwise, Pos_i is supposed to carry surprising information. In order to measure the distance between the posterior and prior distribution, we adopt the Kullback-Leibler divergence as in [15]. So we define the motion-saliency value of Pos_i as below:

$$\begin{aligned} & Surprise(Pos_i, MH(t)) \\ & \quad = KL(P(MH(t) \mid Pos_i), P(MH(t))) \\ & \quad = P(MH(t) \mid Pos_i) \log \frac{P(MH(t) \mid Pos_i)}{P(MH(t))} \end{aligned} \qquad (7)$$

Here we use $S(Pos_i, t)$ to represent the motion-saliency value of the i-th moving object at time t. It also indicates that we have the motion-saliency value for all the points of object trajectories. We then determine the point on the motion trajectory with the highest motion-saliency value at time t as the simulated gaze position $SG(t)$. When there is only one moving object on screen, the simulated gaze trajectory is the same as the object trajectory. In case of multiple moving objects, the simulated gaze position may not stay at a single object trajectory; instead it jumps among different object trajectories. These simulated gaze positions are defined as motion critical positions. Also, the simulated attention shift is defined as the shift among these motion-critical positions.

However, the raw attention shift without any other processing shows a lot of saccadic shifts, or pulse, among object trajectories, so we present here a saccadic shift filter to remove these unrealistically fast attention shifts. Using the motion-saliency module, we have a simulated attention shift sequence $\{SAS(t)\}_{t=1}^{T}$ with $SAS(t)$ denoting the index of the object whose location is the motion critical location at time t. We scan this sequence and remove the saccadic shift in it by counting the time intervals of each value that the SAS sequence repeats. Typically, if $SAS(t)$ equals to $SAS(t+Th)$, where Th is a small integer indicating the threshold for saccadic gaze shift, but the values between them are different, we replace these values by $SAS(t+Th)$, that is,

$$SAS(k) = SAS(t+Th), k = t+1, \ldots t+Th-1 \qquad (8)$$

Experimental results show that the saccadic shift filter is effective in removing the pulse shift in the simulated attention shift sequence and produces more realistic gaze shift patterns.

3. EXPERIMENTAL RESULTS

We test the MCAS model for virtual gazing on a surveillance video clip lasting 27 seconds, which is built from PETS 2001 image sequence. The frame size of the test video is 768*576, with 25 fps. Using the robust motion detector in section 2.1, we obtain the trajectories of three moving objects in the video. Also, we ask 20 people to view this video clip and record 20 gaze data samples from them, using the eye tracker Quick Glance 2 [8] with a sampling rate of 25 Hz. Figure 2 shows three snapshots of this video clip, and Figure 3 shows the 3D view of four gaze trajectories in dashed black line and the three object trajectories marked in green, red and blue respectively. Visually, we can find some common patterns among the gaze trajectories of different people qualitatively, e.g., people are always attracted by the new coming objects, and their gaze shifts between the objects that they try to track. We show that the MCAS model successfully accounts for these observations quantitatively below.

First, we use the motion-saliency model to compute the motion-saliency values for all the points in these three trajectories, shown in Figure 4. After that, we can obtain the critical motion locations by choosing the locations with the highest motion-saliency value every time. However, the raw SAS sequence shows 23 simulated gaze shifts. Applying the saccadic shift filter on the SAS sequence, we remove 15 saccadic gaze shifts with the threshold $Th = 4$. The processed SAS sequence simulates the realistic gaze shift well. Figure 5 shows the comparison between the trajectory of the SAS sequence and eight real gaze trajectories. The trajectory of the SAS sequence is formed by linking the corresponding motion critical positions in the object trajectories, which is also our simulated gaze trajectory. Note that the shift patterns of the two trajectories match well.

From Figure 4 and Figure 5, it is easily verified that when an object come into human vision for the first time, it always has the highest motion saliency value at that time; thus the human gaze is attracted by it, which is also consistent with the results of our observation in Figure 3. Furthermore, a quantitative measure shows that the average number of shifts of the 20 gaze samples is 7.4, while the number of our simulated attention shifts is 8. Moreover, within all the 148 human gaze shifts, 70% of them have the same occurrence time as the simulated attention shifts we have, using a time error of 1 second, which demonstrates the effectiveness of our method for virtual gazing in video surveillance.

4. CONCLUSION

Virtual gazing aims to simulate human gaze shift among multiple moving objects. It creates a ‘Surreal Media’ that a multimedia can be annotated with human attention. In our study, we use motion segmentation to detect moving objects and use the motion-context attention shift (MCAS) model to simulate the attention shifts. Our eye-tracking experiments show that the virtual gazing model can mimic the saccadic trajectory in the multi-object tracking task. We anticipate that it has potential allocations in surveillance video analytics and quantitative measurement of human performance.

5. REFERENCES

[1] M.A. Just and P.A. Carpenter. A theory of reading: From eye fixations to comprehension. Psychological Review, vol. 87(4), pp. 329–354, 1980.

[2] D. Beymer and D. M. Russell. Web gaze analyzer: a system for capturing and analyzing web reading behavior using eye gaze. In Proceedings of CHI’05, 2005, pp. 1913–1916.

[3] R. J.K Jacob. The use of eye movements in human computer interaction techniques: Toward non-command interfaces.

ACM Transactions on Information Systems, vol. 9 (3), pp. 152–169, 1991.

[4] H. Takagi. Development of an eye-movement enhanced translation support system. In Proc. Asian-Pacific Computer Human Interaction Conference (APCHI), 1998, pp. 114–119.

[5] J.L. Sibert, M. Gokturk, and R.A. Lavine. The reading assistant: Eye gaze triggered auditory prompting for reading remediation. In Proceedings of CHI'07, 2000, pp. 101–107.

[6] S.T. Iqbal and B. P. Bailey. Understanding and developing models for detecting and differentiating breakpoints during interactive tasks. In Proceedings of CHI'07, 2007.

[7] Yan Liu, Pei-Yun Hsueh, Lai, J., Sangin, M., Nussli, M.-A., Dillenbourg, P. Who is the expert? Analyzing gaze data to predict expertise level in collaborative applications. Proc. of IEEE International Conference on Multimedia and Expo, 2009.

[8] http://www.eyetechds.com/.

[9] Wang Jian. Integration model of eye-gaze, voice and manual response in multimodal user interface. Journal of Computer Science and Technology, vol.11(5), pp. 512-518, 1996

[10] Helena Grillon, Daniel Thalmann. Simulating gaze attention behaviors for crowds. Journal of Visualization and Computer Animation 20(2-3): 111-119 (2009).

[11] Catherine Pelachaud, Massimo Bilvi: Modelling Gaze Behaviour for Conversational Agents. IVA 2003: 93-100.

[12] L. Li, W. Huang, I.Y.H. Gu, Q. Tian, Foreground object detection from videos containing complex background, ACM Multimedia, 2003.

[13] D. Comaniciu, V. Ramesh, P. Meer, Real-time tracking of non-rigid objects using mean shift, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'00), South Carolina, 2000, pp. 142-149.

[14] L. Itti, P. Baldi, A Principled Approach to Detecting Surprising Events in Video, In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 631-637, Jun 2005.

[15] L. Itti, P. Baldi, Bayesian Surprise Attracts Human Attention, In Advances in Neural Information Processing Systems, Vol. 19 (NIPS*2005), pp. 1-8, Cambridge, MA: MIT Press, 2006.

[16] W. Einhaeuser, T. N. Mundhenk, P. Baldi, C. Koch, L. Itti, A bottom-up model of spatial attention predicts human error patterns in rapid scene recognition, Journal of Vision, Vol. 7, No. 10, pp. 1-13, Jul 2007.

[17] A.L. Yarbus, Eye Movements and Vision. New York: Plenum Press, 1967.

[18] A.T. Duchowski, Eye Tracking Methodologies: Theory and Practice. Springer, 2002.

[19] http://www.youtube.com/watch?v=47LCLoidJh4.

[20] M. Yang, J. Yuan and Y. Wu. Spatial selection for attentional visual tracking, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1-8, 2007.

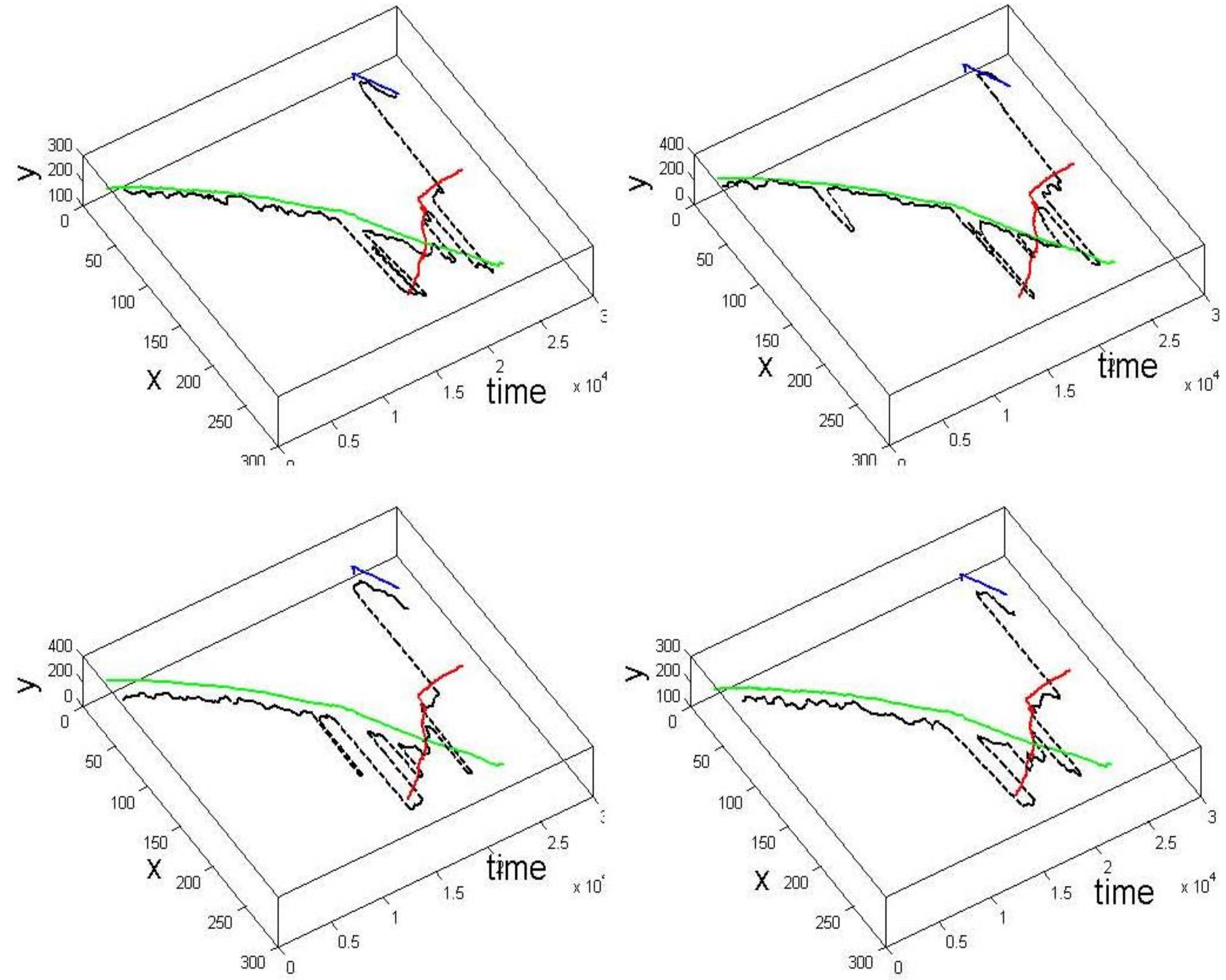

Figure 3. Illustration of the trajectories of the three moving objects in the video marked in green, red and blue respectively, and four gaze trajectories marked with dashed black lines. The four figures only differ in the gaze trajectories, from which we can observe apparent common patterns shared by different gaze trajectories.

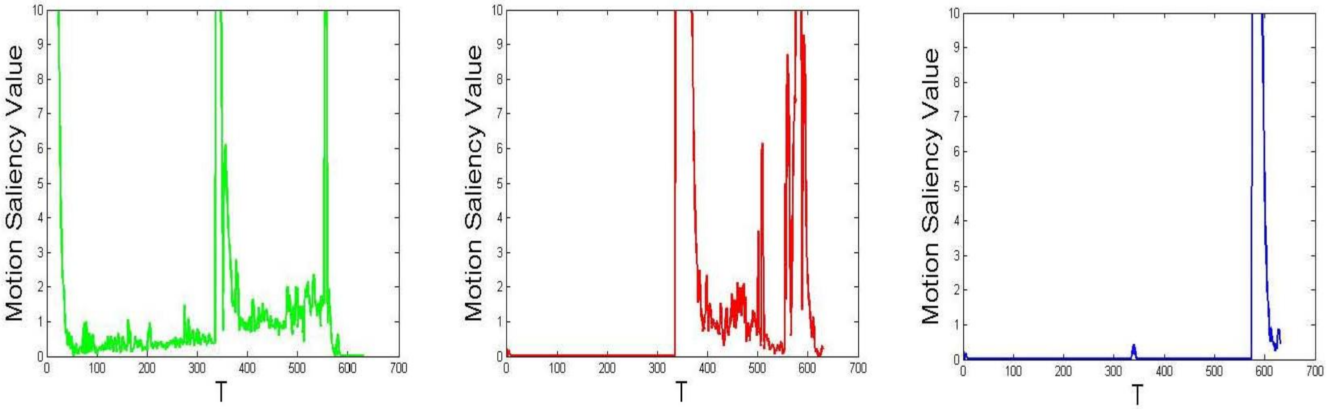

Figure 4. The motion saliency value of the three moving objects marked in green, red and blue in Figure 3

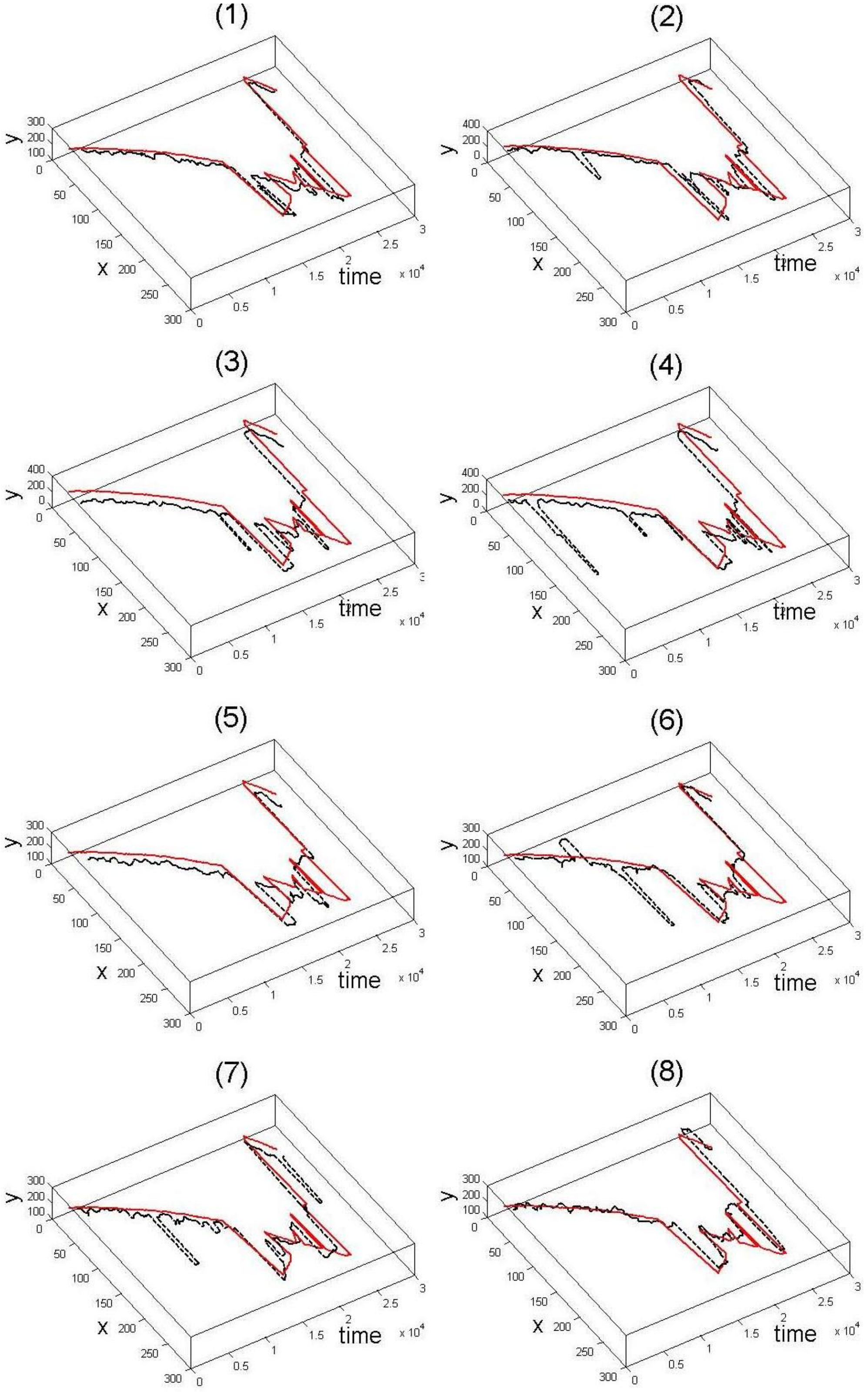

Figure 5. Comparison between our simulated gaze trajectory marked with a solid red line and eight real gaze trajectories marked with dashed black lines.