SALIENT MOTION DETECTION THROUGH STATE CONTROLLABILITY

Karthik Muthuswamy and Deepu Rajan

School of Computer Engineering
Nanyang Technological University
Singapore

ABSTRACT

Salient motion detection is a challenging task especially when the motion is obscured by dynamic background motion. Salient motion is characterized by its consistency while the non-salient background motion typically consists of dynamic motion such as fog, waves, fire etc. In this paper, we present a novel framework for identifying salient motion by modelling the video sequence as a linear dynamic system and using controllability of states to estimate salient motion. The proposed saliency detection algorithm is tested on a challenging benchmark video dataset and the performance is compared with other state-of-the-art algorithms. The results of the comparison indicate that the proposed algorithm demonstrates superior performance when compared to other state-of-the-art methods and with higher computational efficiency.

Index Terms— motion saliency, controllability, dynamic textures

1. INTRODUCTION

Detection of salient motion in a video is useful to support applications like video surveillance, compression and retargeting. It allows fast zooming into areas of suspicion in surveillance videos, higher compression ratio in frames or regions that do not have salient motion and finally, regions without salient motion can be intelligently removed from the frame to resize it for display on devices with different form factors.

Videos shot in natural settings typically consist of backgrounds with dynamic motion and foregrounds with consistent motion. Some videos are shot with a static camera and the objects remain in the frame for a reasonable length of time in the video, e.g. a video of a concert. However, others are shot to track an object of interest. Videos of this type consists of large background motion while having relatively less foreground motion. *Salient motion* is characterized by the consistency in motion irrespective of the intent with which the video is shot. This broad definition of salient motion eliminates background motion like swaying of the leaves etc.

The task of identifying salient motion in videos is a challenging one owing to the various types of motion present in videos. Bugeau et. al. [1] compensate for camera motion and detect groups of pixels arising from consistent motion. However, this method computes consistency from motion only over immediate frames and may not perform well under background motion across a short duration. [2] perform background subtraction by using an adaptive algorithm modelled based on Gaussian mixture probability density, the parameters of which are recursively updated. Frame to frame optical flow measure is utilized to compute the saliency of each pixel as the straight line distance travelled by the pixel across a set of frames in [3]. Mahadevan et.al. [4] proposed a spatio-temporal saliency detection algorithm which used biologically motivated discriminant center-surround saliency hypothesis. It utilizes the dynamic texture model proposed by Doretto et. al. [5] for representing the dynamic motion present in the scene. The saliency measure of a particular location is calculated as the KL divergence between the dynamic texture models of the center and the surround windows. The efficiency of this algorithm however, depends mainly on the size of the center and surround windows. A bottom-up computational framework using low-level features is proposed for identifying visual attention from videos in [11]. In [6], Monnet et. al. use an on-line auto-regressive model to predict the dynamic background motion. Salient motion is detected by comparing the predicted dynamic background and the observed frame. Most recently, Gopalakrishnan et. al. [7] proposed a salient motion detection method by relating saliency to observability of states when the video is modeled as a dynamic texture. This approach requires the observability measure of a frame from different frame buffers in order to identify the final salient motion. However, the number of frame buffers required and the buffer size used to detect salient motion adds a lot of computational overhead.

In this paper, we represent the frames of the video using the dynamic textures model and relate the states of the linear system to saliency using the notion of controllability. In this framework, we exploit the controllability of the dynamic background in estimating the salient foreground motion. The proposed method is generic, in the sense that it generates an accurate saliency map for videos with both local and global motion. This paper is organized as follows. Section 2 discusses the dynamic textures model and the relationship be-

tween salient motion and controllability of a linear system. Section 3 discusses the proposed method for calculating the saliency map and section 4 discusses the quantitative results of the proposed method with state-of-the-art methods before concluding in section 5.

2. STATE-SPACE MODEL TO ESTIMATE SALIENCY

The dynamic texture model [5] represents the observed output of a linear dynamic system $y(t) \in \mathbb{R}^m$ at time t as an ARMA model given by

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Dv(t) \end{aligned} \tag{1}$$

where $A \in \mathbb{R}^{n \times n}$ is the state transition matrix describing the evolution of the state vector $x(t) \in \mathbb{R}^n$. The input Gaussian observation noise is given by $u(t)$ which is an IID realization from the density $\mathcal{N}(0, Q)$, where $Q \in \mathbb{R}^{n \times n}$ is the covariance matrix of the zero-mean gaussian process. $v(t)$ is the output observation noise sampled from $\mathcal{N}(0, R)$, where $R \in \mathbb{R}^{m \times m}$ is the covariance matrix. The matrix $C \in \mathbb{R}^{m \times n}$ relates the current state to the observed output and is called the appearance matrix. Matrix B is the control matrix that determines how the system input affects the state change while matrix D is the feed-through matrix which combines the non-dynamic transfer responses between the input and the output. The matrices A, C, Q and R are estimated using the method proposed by [5] from the measurements $y(1), y(2) \cdots y(\tau)$. A video can hence be completely described using the linear system representation shown above.

2.1. Controllability of a State Space Model

Controllability describes the possibility of finding an input signal that could drive the system from one arbitrary state to another in a finite duration [8]. An LTI system is controllable if and only if its controllability matrix ζ given by $\begin{bmatrix} B & AB & A^2B & \cdots & A^{p-1}B \end{bmatrix}$ is full rank. If $Rank(\zeta) < p$, only a subspace p_c is controllable. Under the given system representation, the inputs will not be able to change the states of $p - p_c$ variables. The above stated definition for the controllability of a linear system provides a subspace of the system which is controllable and another which is not controllable. We explore the relationship between the controllability of the states of a system and salient motion present in a video in the next section.

2.2. Measure of Controllability and its Relation to Saliency

Backgrounds in videos shot with static cameras could be either static as in a highway surveillance video or dynamic due to fluttering of leaves or the motion of waves etc. However, the background in a tracking shot would always be dynamic, e.g. camera tracking a basketball player. Here, we use the notion that salient motion is consistent but not restricted to be in a straight line as described in [3]. The inconsistent background is better modeled using dynamic textures and hence, the inputs can control the states corresponding to the background better than those corresponding to the foreground salient region. Moreover, irrespective of the type of motion present in videos, the consistency in salient motion makes the background motion to be relatively much more controllable than that of the salient motion in a frame buffer. Since the observation $y(t)$ depends on an affine transformation of the state $x(t)$, the controllable states generate an estimated output which is close to the background of the observation. Since salient motion is associated with compact motion in the foreground, the framework allows identification of such motion. Tarokh [9] proposed a quantitative measure for controllability of a linear system based on the eigenvalues of the state transition matrix A. The controllability measure for the i^{th} eigenvalue λ_i is calculated as

$$m_{ci} = |\delta_i|[f_i^* BB^T f_i]^{1/2} \tag{2}$$

where f_i is the left eigenvector corresponding to λ_i of the state transition matrix A and δ_i is $\prod_{j=1}^{n}(\lambda_i - \lambda_j)$ for $i \neq j$. m_{ci} measures the controllability of the eigenvalues of A. However, we are interested in estimating the controllability of the states $x(t)$ of the system. For this, first we evaluate $n_{ci} = f_i^T B$, which is the measure of controllability of the i^{th} state corresponding to the i^{th} eigenvalue of A. The cumulative controllability of the state $x(t)$ is the sum of the controllability measures corresponding to all the eigenvalues, i.e., $n_c(t) = \sum_{j=1}^{n} n_{cj}$. Thus, each state $x(t)$ has its associated controllability vector $n_c(t)$. Finally, a state controlled output estimate $y_c(t)$ of the linear system can be derived in a manner similar to eqn. 1, but using controlled states which are derived by scaling the state vector with the controllability vector as $x_c(t) = n_c(t) * x(t)$, where * indicates component-wise multiplication.

We illustrate the state-controlled estimate of the output in fig. 1 which shows a frame from the original video in fig. 1(a) and the estimate $y_c(t)$ in fig. 1(b). The images are mean subtracted versions of $y_c(t)$ in order to show the effect of controlling the states of the system. Row 1 shows 'Freeway' video, which essentially consists of static but foggy background. The estimated output shows that the controllability of the background is better than that of the foreground regions of moving vehicles, which correspond to large errors in estimation. The second row is the 'Bottle' video of a bottle floating on the waves and hence, has motion in the background. The motion in the waves have been effectively modeled while the consistent motion present in the bottle is estimated poorly. The third row shows the 'Jump' vidoe which is a challenging example having three components - static background, dynamic background due to smoke and salient foreground motion of the cyclist. Here, the estimation of the static background is perfect, while the estimation of the dy-

namic background is poorer. However, the estimation error for salient foreground is worse. Since the controllability of the states corresponding to the salient motion is worse than for the background states, it helps us to identify salient motion. Object appearance encoded by the appearance matrix C is unaffected for those states which cannot be controlled. This serves as a cue for identifying the salient objects in the scene. The advantage of using the proposed method is that the background motion can be estimated using a very small frame buffer τ, as the controllability measure calculates the relative controllability of the matrix A.

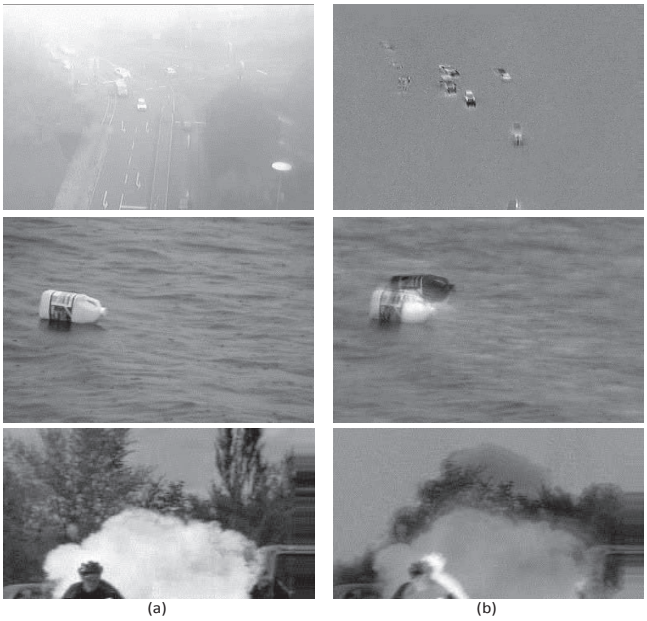

Fig. 1. (a) Observed output $y(t)$ (b) State-controlled estimate $y_c(t)$

3. SALIENCY MAP

The saliency map corresponding to a frame in the video is a gray scale image with higher intensity indicating pixels with higher saliency. We generate the saliency map from two separate maps - the pixel distance map which is suitable for videos shot with a static camera and the sharpness map which is suitable for videos tracking an object. Thus, the two maps serve to complement each other so that the proposed framework caters to videos with generic motion.

3.1. Pixel Distance Map, D_p

The dynamic background motion can be estimated much more accurately than the consistent salient motion as the states of the pixels in the background are relatively more controllable. Hence, the pixel distance between the background pixels in the two frames $y_c(t)$ and $y(t)$ is much lower than that of the salient motion. The pixel distance map $D_p(t)$ for frame t is calculated as $D_p(t) = (y(t) - y_c(t))^2$. Fig. 2 shows the result of applying the proposed algorithm on four different videos namely 'Birds' (row 1), 'Hockey' (row 2), 'Ocean' (row 3) and 'Skiing' (row 4). Fig. 2(a) shows the observed output $y(t)$ while the distance maps $D_p(t)$ generated for the videos are shown in fig. 2(b). It can be seen that in videos having no global motion (Birds, Ocean and Skiing), the controllability of the background provides a reasonable estimate of the dynamic motion as shown by their respective distance maps to $y(t)$. However, the Hockey video consists of highly competing motion which manifests itself in the state-controlled observations. Hence, the $y_c(t)$ for the competing motion has a large distance to the observed $y(t)$.

3.2. Sharpness Map, S_p

Videos that track an object of interest consist of dynamic background motion as well as competing foreground motion. As we are interested in a measure for the controllability of the states of the linear system, motion present in competing objects would often have a controllability measure similar to that of the salient object. The controllability vector does not alter the states of the salient foreground object in the state-controlled output $y_c(t)$ whereas the background appearance is suppressed. As shown in fig. 1, the dynamic texture model can roughly estimate the appearance of the salient object. For a tracking shot, the appearance of the salient object is consistent across frames while the background appearance is not. The sharpness map captures the appearance of the salient object and is estimated by passing $y_c(t)$ through a low-pass filter and comparing the intensity variations between neighbouring pixels in the low-pass filtered version and $y_c(t)$ [10]. The sharpness maps are shown in fig. 2(c). It can be clearly seen that the pixel distance map D_p remove stray salient regions in videos with local motion while the sharpness map S_p removes the stray salient regions in video sequences with dynamic background motion and global motion. The final importance map I_p is calculated as the product of the two normalized maps. The generated final importance map is shown in fig. 2(d).

4. EXPERIMENTAL RESULTS

We evaluate our algorithm for salient motion detection using a benchmark dataset provided by [4]. The database has a total of 18 videos out of which 13 have local motion and 5 have global motion and all of them have competing background motion. Our experiments were conducted using a fixed number of states $n = 10$ and buffer size of 11 frames, which is the minimum that is required to estimate the linear system with 10 states. The sharpness map is estimated for each pixel with a window size of 8×8. Each frame is resized to 128×128 for faster processing. Table 1 shows the Equal Error Rate (EER) as a quantitative evaluation measure

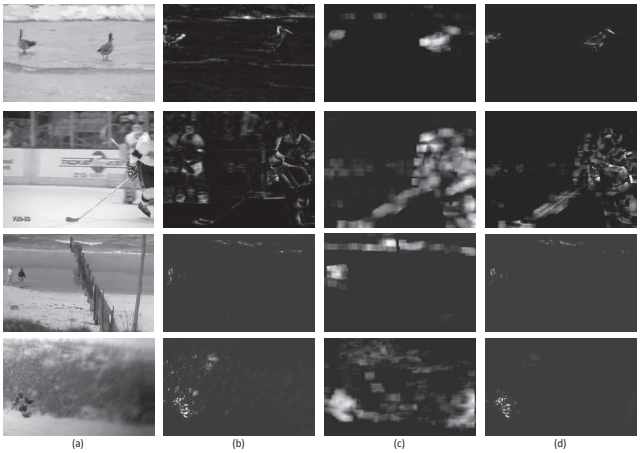

Fig. 2. (a) Observed output $y(t)$ (b) Pixel distance map (c) Sharpness map (d) Final saliency map

to compare our proposed algorithm (SC) with other state-of-the-art algorithms viz., Discriminant Saliency (DS)[4], Monnet et. al (MO)[6], Itti et. al. (IT)[11], Modified Gaussian Mixture Model (GMM)[2] and Sustained Observability (SO) [7]. Only EERs for four videos from each type of motion are shown due to lack of space. The proposed method achieves the second best average EER after [4] for both local as well as global motion. However, it has a computational advantage over the other methods since it uses a smaller frame buffer because the framework is based on a relative measure of controllability of the salient and the background states. Specifically, the best improvement in local motion is for the 'Birds' video which has waves in the background competing with the motion of birds in the foreground. Here, the background is successfully modeled by the dynamic texture and the proposed algorithm ensures that it is more controllable than the foreground region of the birds. Similarly, the best improvement for global motion is for the 'hockey' video. Overall, the proposed method performs the best in 9 out of the 18 videos with 6 out of 13 for local motion and 3 out of 5 for global motion.

Table 1. Comparison of EER measures (in %)

	Other Motion saliency algorithms					
Motion type	DS	MO	IT	GMM	SO	SC
Local Motion						
-Birds	5	7	19	23	9.2	3
-Freeway	6	31	43	25	8.7	9.2
-Bottle	2	17	5	25	3.9	2.1
-Jump	15	23	25	39	22	13.8
Global Motion						
-Cyclists	8	28	41	36	17.9	8.8
-Hockey	24	29	28	39	27	18.1
-Surf	4	10	30	23	7.6	7.5
-Surfers	7	10	24	35	9	5.2
Avg. EER (Loc.)	6.8	14.9	25.4	28.9	11.4	8.42
Avg. EER (Glo.)	9.2	18.8	30.8	34.6	17.7	11.2
Avg. EER	7.6	16	26.2	29.7	12.6	9.2

5. CONCLUSION

We propose a novel approach to detect saliency from motion using the concept of controllability. The proposed method performs second best among state-of-the-art methods while having a computationally efficient approach. The proposed approach has been shown to be robust enough to handle videos with generic motion types.

6. REFERENCES

[1] A. Bugeau and P. Prez, "Detection and segmentation of moving objects in complex scenes," *CVIU*, vol. 113, pp. 459–476, 2009.

[2] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," 2004, vol. 2, pp. 28–31 Vol.2.

[3] L. Wixson, "Detecting salient motion by accumulating directionally-consistent flow," *IEEE Transactions on PAMI*, vol. 22, pp. 774–780, 2000.

[4] Vijay Mahadevan and Nuno Vasconcelos, "Spatiotemporal saliency in dynamic scenes," *IEEE Transactions on PAMI*, vol. 32, pp. 171–177, 2010.

[5] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto, "Dynamic textures," *IJCV*, vol. 51, no. 2, pp. 91–109, 2003.

[6] A. Monnet, A. Mittal, N. Paragios, and V. Ramesh, "Background modeling and subtraction of dynamic scenes," in *Proceedings of the Ninth IEEE ICCV*, 2003.

[7] V. Gopalakrishnan, Y. Hu, and D. Rajan, "Sustained observability for salient motion detection," in *Proceedings of the 10th ACCV*, 2011.

[8] P. Albertos and A. Sala, "Multivariable control systems - an engineering approach," *Automatica*, pp. 65–68, 2005.

[9] M. Tarokh, "Measures for controllability, observability and fixed modes," *IEEE Transactions on Automatic Control*, vol. 37, pp. 1268–1273, 1992.

[10] F. Crete, T. Dolmiere, P. Ladret, and M. Nicolas, "The blur effect: perception and estimation with a new no-reference perceptual blur metric," in *SPIE*, 2007.

[11] L Itti and P. Baldi, "A principled approach to detecting surprising events in video," in *CVPR*, 2005, pp. 631–637.