Motion-Based Generator Model: Unsupervised Disentanglement of Appearance, Trackable and Intrackable Motions in Dynamic Patterns

Jianwen Xie[1,*], Ruiqi Gao[2,*], Zilong Zheng[2], Song-Chun Zhu[2], Ying Nian Wu[2] (*equal contribution)

[1] Hikvision Research Institute, [2] University of California, Los Angeles

Abstract

Understanding dynamic patterns requires a disentangled representational model that separates the factorial components. A commonly used model for dynamic patterns is the state space model, where the state evolves over time according to a transition model and the state generates the observed image frames according to an emission model. To model the motions explicitly, it is natural for the model to be based on the motions or the displacement fields of the pixels. Thus in the emission model, we let the hidden state generate the displacement field, which warps the trackable component in the previous image frame to generate the next frame while adding a simultaneously emitted residual image to account for the change that cannot be explained by the deformation. The warping of the previous image is about the trackable part of the change of image frame, while the residual image is about the intrackable part of the image. We use a maximum likelihood to learn the model parameters that iterates between inferring latent noise vectors that drive the transition model and updating the parameters given the inferred latent vectors. Meanwhile we adopt a regularization term to penalize the norms of the residual images to encourage the model to explain the change of image frames by trackable motion. Unlike existing methods on dynamic patterns, we learn our model in unsupervised setting without ground truth displacement fields or optical flows. In addition, our model defines a notion of intrackability by the separation of warped component and residual component in each image frame. We show that our method can synthesize realistic dynamic pattern, and disentangling appearance, trackable and intrackable motions. The learned models can be useful for motion transfer, and it is natural to adopt it to define and measure intrackability of a dynamic pattern.

Motion-based generator model

Let $\mathbf{I} = (\mathbf{I}_t, t = 0, 1, ..., T)$ be the observed video sequence of dynamic pattern, where $\mathbf{I}_t$ is a frame at time t. The motion-based model for the dynamic patterns consists of:

$$s_t = (s_t^M, s_t^R) = f_1(s_{t-1}, h_t), \tag{1}$$
$$M_t = (\delta(x, y), \forall (x, y) \in D) = f_2(s_t^M), \tag{2}$$
$$R_t = f_3(s_t^R), \tag{3}$$
$$I_t = f_4(I_{t-1}, M_t), \tag{4}$$
$$\mathbf{I}_t = I_t + R_t + \epsilon_t, \tag{5}$$

where $f = (f_i, i = 0, 1, 2, 3)$ are neural networks parameterized by $\theta = (\theta_i, i = 0, 1, 2, 3)$.

(i) Equation (1) is the transition model, where s_t is the state vector, h_t is a hidden Gaussian white noise vector. f_1 defines the transition from s_{t-1} to s_t. The state vector s_t consists of two sub-vectors. One is s_t^M for motion. The other is s_t^R for residual.
(ii) In Equation (2), s_t^M generates the motion M_t of the trackable part I_{t-1} of the image frame $\mathbf{I}_{t-1}$, where M_t is the field of pixel displacement, which consists of the displacement $\delta(x, y)$ of pixel (x, y). f_2 defines the mapping from s_t^M to M_t.
(iii) In Equation (4), M_t is used to warp the trackable part I_{t-1} of the image frame $\mathbf{I}_{t-1}$ by a warping function f_4, which is given by bilinear interpolation. No unknown parameter in f_4.
(iv) In Equation (3), s_t^R generates residual image R_t. f_3 defines the mapping from s_t^R to R_t.
(v) In Equation (5), the image frame $\mathbf{I}_t$ is the sum of the warped image I_t and the residual image R_t, plus a Gaussian white noise error ϵ_t.
(vi) In Equation (6), the initial trackable frame I_0 is generated by an generator f_0 from an appearance hidden variable c that follows Gaussian distribution. To initialize $\mathbf{I}_0$, we use:

$$I_0 = f_0(c),\ R_0 = f_3(s_0^R),\ \mathbf{I}_0 = I_0 + R_0 + \epsilon_0. \tag{6}$$

Learning by alternative back-propagation through time

Let $p(h)$ be the Gaussian white noise prior. Let $p_\theta(\mathbf{I}|h) \sim \mathrm{N}(f_\theta(h), \sigma^2 I)$ be the conditional distribution of the video sequence $\mathbf{I}$ given h. The marginal distribution of $\mathbf{I}$ is $p_\theta(\mathbf{I}) = \int p(h) p_\theta(\mathbf{I}|h) dh$ with the latent variable h integrated out.

We estimate the model parameter θ by the maximum likelihood method that maximizes the observed-data log-likelihood $\log p_\theta(\mathbf{I})$. The gradient of the log-likelihood $\log p_\theta(\mathbf{I})$ is:

$$\frac{\partial}{\partial \theta} \log p_\theta(\mathbf{I}) = \frac{1}{p_\theta(\mathbf{I})} \frac{\partial}{\partial \theta} p_\theta(\mathbf{I}) = \mathrm{E}_{p_\theta(h|\mathbf{I})} \left[\frac{\partial}{\partial \theta} \log p_\theta(h, \mathbf{I}) \right], \tag{7}$$

where $p_\theta(h|\mathbf{I}) = p_\theta(h, \mathbf{I}) / p_\theta(\mathbf{I})$ is the posterior distribution of the latent h given the observed X. The above expectation can be approximated by Monte Carlo average.

The learning algorithm iterates the following two steps:
(1) **Inference step**: given the current θ, sample h from $p_\theta(h^{(\tau)}|\mathbf{I})$ by the Langevin dynamics

$$h^{(\tau+1)} = h^{(\tau)} + \frac{\delta^2}{2} \frac{\partial}{\partial h} \log p_\theta(h^{(\tau)}|\mathbf{I}) + \delta N(0, I), \tag{8}$$

(2) **Learning step**: given h, update θ by stochastic gradient descent

$$\Delta \theta \propto \frac{\partial}{\partial \theta} \log p_\theta(h, \mathbf{I}), \tag{9}$$

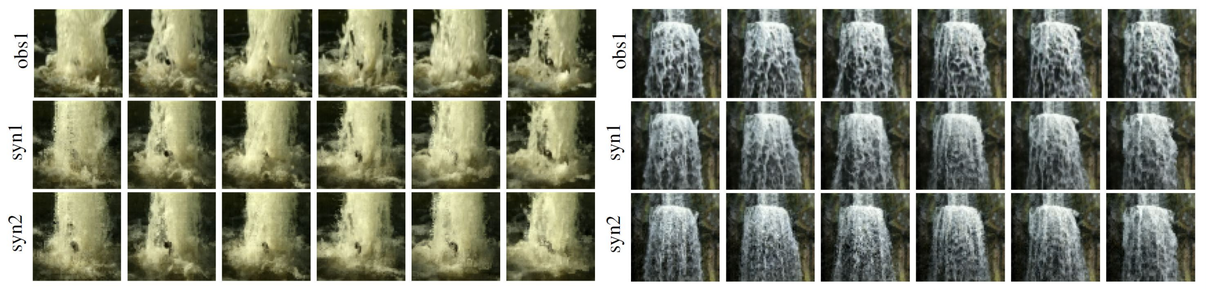
Since $\frac{\partial}{\partial h} \log p_\theta(h|\mathbf{I}) = \frac{\partial}{\partial h} \log p_\theta(h, \mathbf{I})$, both steps involves derivatives of

$$\log p_\theta(h, \mathbf{I}) = -\frac{1}{2} \left[\|h\|^2 + \frac{1}{\sigma^2} \|\mathbf{I} - f_\theta(h)\|^2 \right] + \text{const},$$

where the constant term does not depend on h or θ. Both can be computed by back-propagation through time. To encourage the model to explain the video sequence $\mathbf{I}$ by the trackable motion, we add to the log-likelihood $\log p_\theta(\mathbf{I})$ a penalty term $-\lambda_1 \|R_t\|^2$. To encourage the smoothness of the inferred displacement field M_t, we add another penalty term $-\lambda_2 \|\Delta M_t\|^2$. We estimate θ by gradient ascent on $\log p_\theta(\mathbf{I}) - \lambda_1 \sum_t \|R_t\|^2 - \lambda_2 \sum_t \|\Delta M_t\|^2$.

Experiment 1: Dynamic pattern synthesis

We learn the model for dynamic textures, which are sequences of images of moving scenes that exhibit stationarity in time. We learn a separate model from each example.



(a) fountain (b) waterfall

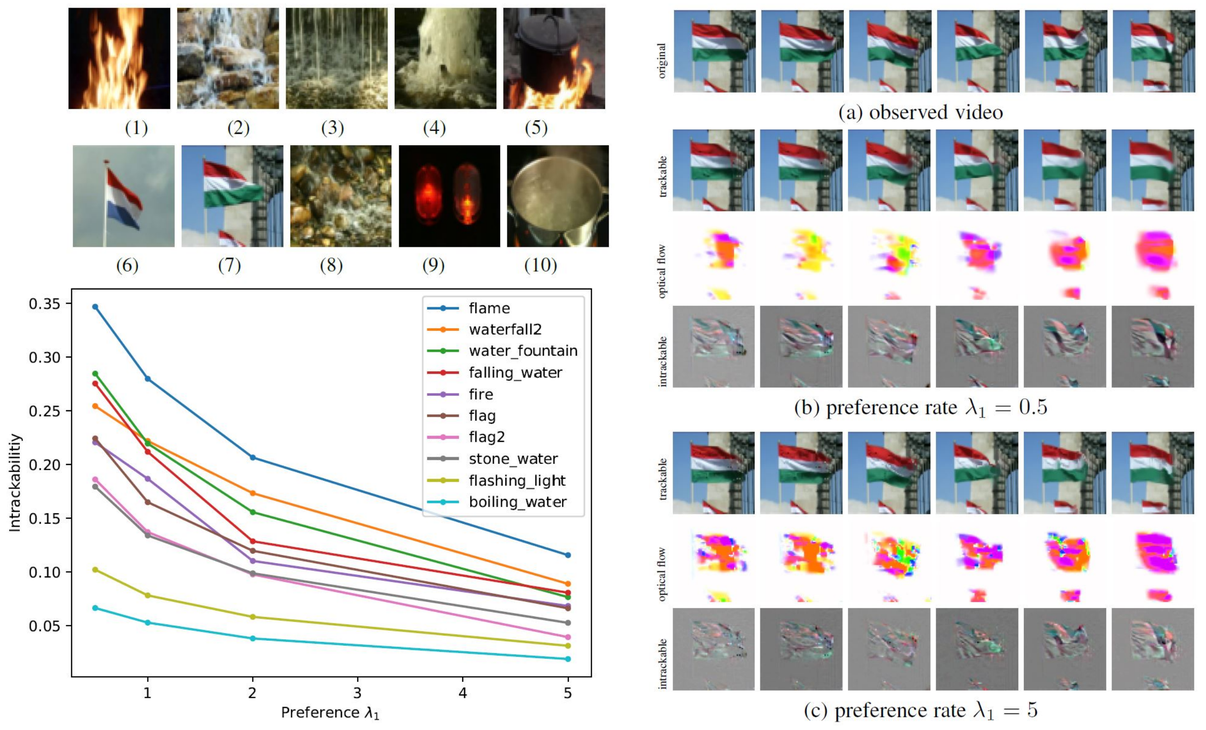
Experiment 2: Unsupervised disentanglement of appearance and motion

We learn the model from only one single video and unsupervisedly disentangle the motion and appearance of the video, and then transfer the motion to the other appearances.

(a) observed facial expression (happiness) (a) observed facial expression (surprise)
(b) learned motion (happiness) (b) learned motion (surprise)
(c) apply the learned motion to some appearances in the testing set (c) apply the learned motion to some appearances in the testing set
(d) apply the learned motion to cartoon appearances (d) apply the learned motion to cartoon appearances
(e) apply the learned motion to animal appearances (e) apply the learned motion to animal appearances

Experiment 3: Disentanglement of trackable and intrackable motions

In the context of our model, we can define intractability as the ratio between the average of ℓ_2 norm of the intrackable residual image R_t and the average of the ℓ_2 norm of the observed image $\mathbf{I}_t$. This ratio depends on the penalty parameter λ_1 of the ℓ_2 norm of R_t used in the learning stage. This penalty parameter corresponds to the subjective preference. The larger the preference λ_1 is, the larger extent to which we interpret a video by trackable contents, the less the residuals, and the less intrackability score.



(a) observed video
(b) preference rate $\lambda_1 = 0.5$
(c) preference rate $\lambda_1 = 5$

Appendix

Left: An illustration of the framework of the proposed model-based generator model.
Right: Visualization of displacement field.