Future Qualitative Activity Graph Prediction

Alexia Toumpa[1] **and Anthony G. Cohn**[1, 2, 3]

[1] School of Computing, University of Leeds
[2] School of Control and Engineering, Shandong University, Jinan, 250061, China
[3] College of Electronic and Information Engineering, Tongji University, China
A.Toumpa@leeds.ac.uk, A.G.Cohn@leeds.ac.uk

Abstract

Interaction and action anticipation remains a challenging problem, especially considering the generalizability constraints of trained models from visual data or exploiting visual video embeddings. To overcome these constraints, we present an initial investigation of a novel approach for solving the task of interaction anticipation between objects in a video scene by utilizing a qualitative spatial graph representation. A convolutional recurrent neural network architecture learns in a self-supervised way to predict qualitative spatial graph structures of future object interactions, while being decoupled from visual information.

Introduction

Performing long-range predictions of spatio-temporal information from video data is a challenging problem, evident in many real-world applications, such as self-driving, robot control, and human-robot collaboration, as well as perception tasks, as action prediction and object tracking.

Some prior works focused on *short-term video prediction*, such as the ContextVP network (Byeon et al. 2018) which models contextual dependencies. Furthermore, the MCnet network (Villegas et al. 2017) acts on the motion and content of the video data, separating them into different encoder paths and performing prediction of future frames considering the observed motion. Also, inspired by ‘predictive coding’, frame predictions with PredNet (Lotter, Kreiman, and Cox 2016) are based on the deviations of local predictions from every layer of the architecture.

Other works have focused on *long-range prediction* networks, such as the Convolutional LSTM network (Xingjian et al. 2015), which integrates convolutions into state transitions of a recurrent neural network. Also, action-conditioned video prediction networks (Finn, Goodfellow, and Levine 2016) model pixel motion for learning physical object motion. Another long-range prediction network proposes a memory transition mechanism which memorizes local appearance and motion for short-term spatio-temporal predictions and exploits an attention mechanism on previous memory cells for long-range predictions (Wang et al. 2018). Similar to the MCnet, the DRNet (Denton et al. 2017) considers two separate Encoder pathways for the object pose and the content of the video for better prediction quality. Moreover, the VPN network (Kalchbrenner et al. 2017) models the factorization of the joint likelihood of the video by estimating local dependencies of neighboring pixels. Also, two-stream recurrent neural networks (Xu et al. 2018) are employed for capturing the different frequency domain information.

However, pixel-level predictions of future video frames have high uncertainty after a few frames causing blurriness of the output. Moreover, learning from visual features constrains the model’s generalizability across different domains as the features learned are based on the visual appearances present in the dataset. In this paper we present an initial investigation of a novel approach for addressing the problem of spatio-temporal anticipation considering object interactions from real-world video data. The proposed approach considers class-agnostic objects and learns high-level features of object interactions. Hence, we obtain information about how the scene is going to change in the future in reference to the activity taking place.

We exploit high-level qualitative graphical structures to represent object interactions present in a video, to abstract from the feature space of the image scene and attain representation generalization across different domains. We learn from these qualitative graphs in a self-supervised way the spatio-temporal correlations between interactions and we predict future qualitative graphs of future object interactions by exploiting a Convolutional LSTM network architecture. Due to the high-level representations of interactions, our method is not video frame dependent and can predict future object interactions with high *Jaccard index*, *Accuracy* and *F1* scores. Figure 1(a) illustrates an overview of the proposed method. The objectives of this work are:

- to create a 3D tensor representation that captures the information from high-level relational graphs;
- to predict future interactions in frames-independent time intervals, considering graph representations of object interactions, based on episode detections.

Interaction Sequence Modeling

As graphical structures are able to capture high level information and achieve generalization across different domains, we exploit qualitative relational graphs to represent activities between objects in a video scene considering their spatio-temporal interactions.

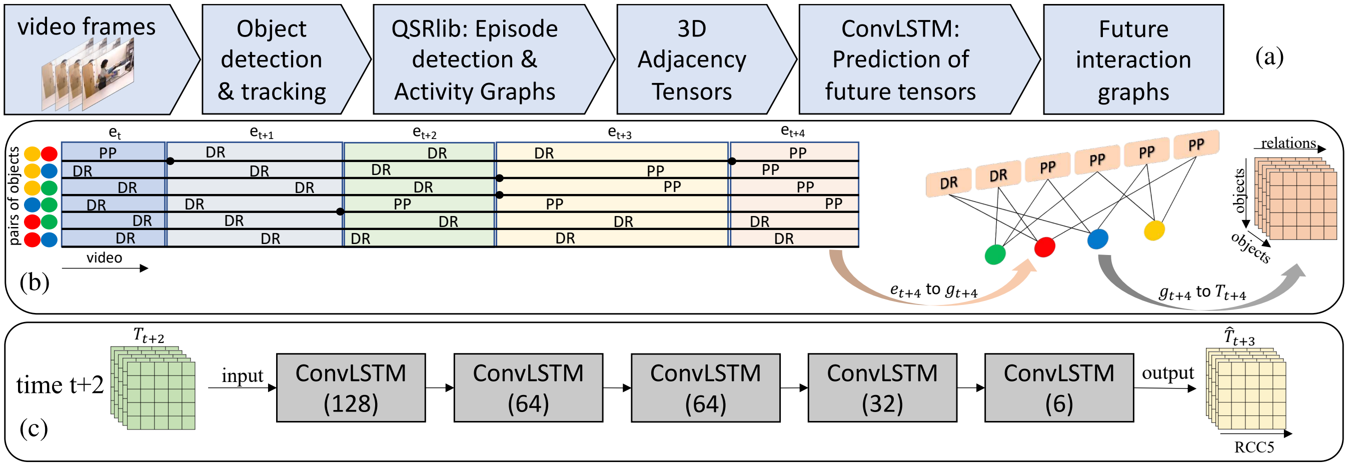

Figure 1: (a) An overview pipeline of the proposed approach for future interaction graph prediction. (b) Episode detection in a demo video with several color-coded objects. For simplicity we visualize only the RCC5 relations. (c) A ConvLSTM network is employed for predicting the future RCC5 relationships from qualitative spatio-temporal tensor representations.

For this purpose, we employ a variant of *Activity Graphs* (Sridhar, Cohn, and Hogg 2010a,b) (vAGs) to represent entity, *i.e.* object, interactions present in a video. These graph representations (g) comprise two layers of vertices where each layer consists of a single type of node and only nodes in adjacent layers can be connected with each other. The bottom (object) layer contains the set of vertices representing the interacting entities of the video, and the top (spatial) layer consists of the vertices with the spatial relations describing the spatial interaction of the entities. The spatial relations we capture are:

- for every pair of objects, the relationships from the *Region Connection Calculus* (RCC5) (Randell, Cui, and Cohn 1992; Cohn et al. 1997) which consist of the relations: ‘discrete’ (DR), ‘partially overlapping’ (PO), ‘proper part’ (PP, PPi), and ‘equal’ (EQ),
- for every pair of objects, the relationships from the *Qualitative Trajectory Calculus* (QTC) (Van de Weghe et al. 2006; Delafontaine, Cohn, and Van de Weghe 2011) which contains the relations: ‘-,-’, ‘-,0’, ‘-,+’, ‘0,-’, ‘0,0’, ‘0,+’, ‘+,-’, ‘+,0’, and ‘+,+’; where the pair ‘α, β’ represents the relative motion of each object towards the other and ‘+’ means motion away, ‘-’ means motion towards, and ‘0’ means no relative motion,
- for every object, a binary state of *Moving or Stationary* (MoS) ,
- and the *Cardinal Direction* of motion (CarDir) (Frank, Mark, and White 1991) for every moving object in the scene, that corresponds to the set of relations: ‘north’ (N), ‘north east’ (NE), ‘east’ (E), ‘south east’ (SE), ‘south’ (S), ‘south west’ (SW), ‘west’ (W), ‘north west’ (NW), and ‘equal’ (EQ).

The maximum period of time throughout which a spatial relation between the video entities occurs, whilst before and after that time a different spatial relation holds, is an *episode* (e), and multiple episodes define the sequence of spatial relations obtained in every interaction (Fig. 1(b)).

Tensor Representation

At training time, for every video we utilize object proposals to define entities and extract a vAG (g), representing each detected episode and consisting of all the objects involved. *E.g.* in Figure 1(b) episode e_{t+4} is represented by vAG g_{t+4}. The temporal information for every vAG is represented by ordering them in temporal episode-detection order.

We exploit a 3D tensor representation for the vAGs (Fig. 1(b)). Each tensor describes the spatial relationships holding in an episode. Hence, a sequence of tensors carries all the interactions present in a video. A vAG tensor $T \in \{0,1\}^{O\times O\times R}$ is based on the construction of a 3D adjacency matrix between all entities and spatial relations of an episode, where O is the number of entities and R the number of relations. Thus, the values in T are assigned based on Equation 1 where o_1, o_2, and r are the locations of the two objects and the location of the relation ($relation_r$) respectively, in T.

$$T[o_1, o_2, r] = \begin{cases} 1 & \text{if } relation_r(obj_{o_1}, obj_{o_2}) = \text{True} \\ 0 & \text{if } relation_r(obj_{o_1}, obj_{o_2}) = \text{False} \end{cases} \tag{1}$$

Since the size of the tensor is static, it doesn’t change depending on the number of detectable objects; some object rows will be filled with zeros if fewer than O objects are detected. A zero value in a detected object specifies that the specific relation between that object and another is not present, whereas a zero value for a non-detected object means that the object is not present. To explicitly differentiate these two cases, for every relational set we add an extra relation ‘not applicable’ (N/A) which applies to all non-detected objects. We select O to be sufficiently big so the number of detected objects does not exceed the tensor’s size.

Convolutional LSTM network for Qualitative Interactions Prediction

As our application requires capturing long-range dependencies in multi-dimensional tensors for representing qualitative spatio-temporal information we exploit Convolutional LSTM networks (ConvLSTM) (Xingjian et al. 2015) that were introduced as an extension of the Fully Connected LSTM network (FC-LSTM) considering convolution structures in the input-to-state and the state-to-state transitions. All features are represented by 3D tensors with dimensions (height × width × channels) and the matrix multiplications are replaced with tensor convolutions. The parameters of a ConvLSTM are the input weights $W_x \in \mathbb{R}^{K\times K\times C}$ and the recurrent weights $W_h, W_o \in \mathbb{R}^{K\times K\times F}$ with K, C, and F denoting the kernel size, the number of channels of the input $X_t \in \mathbb{R}^{H\times W\times C}$ and the hidden states $H_t \in \mathbb{R}^{H\times W\times F}$ respectively. The key equations of ConvLSTM are:

$$\begin{aligned}
i_t &= \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i)\\
f_t &= \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f)\\
C_t &= f_t \circ C_{t-1} + i_t \circ tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c)\\
o_t &= \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_t + b_o)\\
H_t &= o_t \circ tanh(C_t)
\end{aligned} \tag{2}$$

where $\sigma(\cdot)$ represents a sigmoid function, '$*$' denotes a convolution, '$\circ$' the Hadamard product, and $i_t, f_t, o_t, C_t \in \mathbb{R}^{K\times K\times F}$ are the input gate, forget gate, output gate and memory cell respectively. The input to our network is of dimension $(M \times H \times W \times C)$, where M is the number of samples, $H = W = O$ represent the number of detectable objects, and C is the set of the spatial relations captured (29). Also, the output tensor is of dimension $(H \times W \times F)$, where $H = W = O$ and F is the number of relations to be predicted. We selected the output to describe spatial positions of the objects based only on RCC5, since we consider such information as the most useful for many real-world applications. Thus we set $F = 6$, where 6 is the number of RCC5 relations with an additional N/A relationship for non-detected objects.

At deployment time, this network processes incrementally a sequence of vAGs for predicting the future state of objects' interactions. The networks updates its internal recurrent state for every episode of interactions, accumulating in that sense past information and enhancing its future predictions. The complete representation of the input/output tensors and the proposed pipeline are shown in Figure 1(c). The input data capture the information of object interactions for all the relations (RCC5, QTC, MoS, CarDir), though the output represents the spatial interactions from the RCC5 set of relationships only.

Self-supervised The network is trained in a self-supervised way, by exploiting the sequential nature of our tensor data. More specifically, at timestep t the input comprises of the tensor data at time t (T_t) and $t-1$ (T_{t-1}), from which T_{t-1} is used as the model's input and T_t is compared against the predicted output to update the model's weights.

Training Loss Due to the nature of our data, *i.e.* binary sparse tensors, our task is to correctly predict the correct tensor as a multi-class classification problem for every pair of objects between the different spatial relations we capture. Hence, for updating the weights of the network in every iteration we minimize a weighted categorical cross-entropy loss function, defined as:

$$\mathscr{L} = \frac{1}{M}\sum_{k=1}^{K}\sum_{m=k}^{M} w_k \cdot y_m^k log(h_\theta(x_m, k)) \tag{3}$$

where M represents the number of training examples, K is the number of classes, w_k is the weight of class k, h_θ represents the model with neural network weights θ, y_m^k is the target label for the training example m of class k, and x_m denotes the input of the training example m. The weights of each class k (w_k) are set percentage-wise depending on the overall detection of each one across the whole dataset. Hence a relation that appears often will have a low weight, whereas a relation that appears rarely in the dataset's interactions will have a higher weight.

Model Architecture Inspired by the model architecture proposed by (Xingjian et al. 2015), our proposed network comprises of a series of layers of ConvLSTM modules with 128, 64, 64, and 32 hidden states outputting to a 6 channeled tensor (Fig. 1(c)). Hence, the output tensors are of dimension $(O \times O \times 6)$ and the input tensors are of dimension $(O \times O \times 29)$, where O represents the number of detected objects. Furthermore, the kernel size is set to $(1, 1)$ as every value in the tensor is independent of its neighbors, and the weights are initialized based on the LeCun uniform distribution (LeCun et al. 2012).

Training Hyper-parameters During training, the Adam optimizer was employed for the update of the weights. The learning rate started from 0.01 along with a scheduler to reduce the value of the learning rate every 1000 epochs with a factor of 0.1. We trained the model until convergence for a maximum of 2.5k epochs. Moreover, the batch size was set to 5 considering the minimum number of episodes captured from a single video. Thus, no batch disturbs the temporal ordering of the data by shuffling or concatenating data from different videos. We also added a Lasso regularization term of $\lambda|w|$ with $\lambda = 1e-4$ in the loss function to avoid overfitting of the model.

Experiments

Dataset

We trained and evaluated the proposed approach on the CAD-120 dataset (Koppula, Gupta, and Saxena 2013) whilst exploiting the groundtruth bounding boxes of object positions. The CAD-120 dataset comprises of 120 RGB-D sequences of frames of everyday-life activities, capturing human-object interactions in various scenes, *e.g.* office, kitchen, etc. From every video of the dataset, tensor representations of the object interactions are created by considering the input relations while extracted from the QSRlib library (Gatsoulis et al. 2016). Due to the static size of the tensors we set the maximum number of detected objects (O) to be 10 which is adequate for capturing all the object interactions for the employed dataset. Also, in every epoch, the tensor data are shuffled along both the object axes so no correlation between the rows is learned.

Evaluation

We performed experiments using different qualitative spatial information to evaluate how the incorporation of each rela-

Table 1: Quantitative results of the experiments on the test set.

					Training parameters	
Model	J.I. (↑)	W.C.E. (↓)	C.Acc. (↑)	F1 (↑)	num. parameters	batch
B_1: RCC5	0.4228	0.9765	0.9198	0.4504	164,904	4
B_2: RCC5+QTC	0.5747	0.7453	0.9454	0.5921	170,024	5
B_3: RCC5+QTC+CarDir	0.6329	0.8512	0.9469	0.6472	175,144	5
RCC5+QTC+CarDir+MoS	**0.6477**	**0.6133**	**0.9621**	**0.6659**	176,680	5

tional set helps improve the predicted output. The predicted tensors represent the future interactions of the next episode. For baselines B_3 and B_2 we set the batch size to 5 whereas for B_1 to 4, due to the smaller number of episodes detected, fewer spatial relations denote fewer episodes.

Inspired by the evaluation metrics in the *instance segmentation* literature, for quantifying the overlap of 1s in the predicted tensors over 1s in the groundtruth tensors of RCC5 spatial interactions, one of the metrics we employ is the *Jaccard similarity index* (J.I.) (Eq. 4) for multiple classes, which considers the number of classes (K) and the true positives (TP_k), false positives (FP_k) and false negative (FN_k) for every class (k).

$$\mathscr{J}_{mc} = \frac{1}{|K|} \sum_{k=1}^{K} \frac{\text{TP}_k}{\text{FP}_k + \text{FN}_k + \text{TP}_k} \tag{4}$$

Moreover, we report the *categorical accuracy* measure (C.Acc.), the *F1-score* (Van Rijsbergen 1979) (F1), as well as the *weighted cross-entropy* loss value (W.C.E.) for every experiment. Our method was evaluated on 25% of randomly-picked unseen video data and the results in Table 1 demonstrate that the proposed approach achieves the best results in all reported metrics [1]. More specifically, the proposed approach combining the information of 1) the spatial location, 2) the relative motion, 3) the absolute motion, and 4) the direction of the absolute motion of the objects, can achieve an increase of the Jaccard index score of 2%, 13% and 53% compared to the baselines B_3, B_2 and B_1 respectively. Moreover, we acquire an increase of the categorical accuracy of 1.6%, 1.8% and 4.6%, as well as an increase of the F1-score of 2.9%, 12.5% and 47.8% compared to the baselines B_3, B_2 and B_1, respectively. The weighted cross-entropy loss value shows a significant improvement of 27.9%, 17.7% and 37.2% compared to the baselines B_3, B_2 and B_1, respectively, since it considers the imbalance of the data by applying a weight at each relation. Furthermore, these results were attained with a maximum growth of 7.1% in the model size.

Some qualitative results are illustrated in Figure 2 along with the corresponding vAGs, with the pair-wise relations, for one of the interactive objects (object α). Figure 2 illustrates a visual representation of a two dimensional snap shot of the model's input, along with the model's prediction and corresponding groundtruth tensor, of the relationships between all objects and object α. For simplicity we only show the graph information for RCC5 and QTC relations, omitting the MoS and CarDir relations from the input tensor.

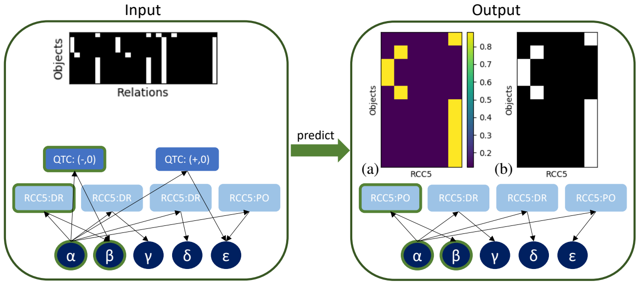

Figure 2: Qualitative results in an example case for the interactions with object α. Output matrix (a) corresponds to the network prediction, whereas matrix (b) represents the groundtruth relations. White cells contain the value 1 and black cells the value 0. Yellow and purple cells are the predicted values closer to 1 and 0, respectively.

It is evident that the motion information (QTC:(-,0)) as well as the direction of motion, signify that object β is moving towards object α. Hence, in the predicted vAG a PO RCC5 relation holds between objects α and β. We binarize the values of the predicted tensors by setting the switch point to 0.5. Thus, values greater than 0.5 are considered as 1.0 and 0.0 otherwise. Hence, by binarizing the predicted tensor our prediction tensor for this example maps exactly to the groundtruth.

Conclusions

We have presented an initial study of a novel approach for solving the task of interaction anticipation whilst exploiting high-level qualitative spatial representations and training a ConvLSTM network in a self-supervised way. Our results demonstrate that exploiting a rich set of high-level relations is a promising direction for predicting future spatial interactions, whilst not being frame dependent.

This is ongoing research, and we are working towards cross-validating the performance of our trained model in various real-world everyday-life activity datasets, as well as investigating the impact of the incorporation of object visual feature embeddings. Moreover, we are working towards evaluating the proposed approach against the works of Srivastava, Mansimov, and Salakhudinov (2015) and Chen et al. (2022). We are also focusing on conducting real-world experiments, in which a robot agent is asked to complete an activity initiated by a human agent, to showcase the impact of such episodes-based predictions in a human-robot collaboration scenario.

[1] ↑ indicates 'highest is best' and ↓ indicates 'lowest is best'

Acknowledgments

This work has been partially supported by an Engineering and Physical Sciences Research Council (EPSRC) Studentship to the first author, by the European Union's Horizon 2020 research and the innovation program under grant agreement No. 825619 (AI4EU), and by an Alan Turing Institute Fellowship to the second author.

References

Byeon, W.; Wang, Q.; Kumar Srivastava, R.; and Koumoutsakos, P. 2018. Contextvp: Fully context-aware video prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Chen, G.; Zhang, W.; Lu, H.; Gao, S.; Wang, Y.; Long, M.; and Yang, X. 2022. Continual Predictive Learning from Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10728–10737.

Cohn, A. G.; Bennett, B.; Gooday, J.; and Gotts, N. M. 1997. Qualitative Spatial Representation and Reasoning with the Region Connection Calculus. *GeoInformatica*.

Delafontaine, M.; Cohn, A. G.; and Van de Weghe, N. 2011. Implementing a Qualitative Calculus to Analyse Moving Point Objects. *Expert Systems with Applications*, 38(5): 5187–5196.

Denton, E. L.; et al. 2017. Unsupervised Learning of Disentangled Rrepresentations from Video. In *Advances in Neural Information Processing Systems*.

Finn, C.; Goodfellow, I.; and Levine, S. 2016. Unsupervised Learning for Physical Interaction Through Video Prediction. In *Advances in Neural Information Processing Systems*.

Frank, A.; Mark, D.; and White, D. 1991. Qualitative Spatial Reasoning About Cardinal Directions. In *Proc. of the 7th Austrian Conf. on Artificial Intelligence. Baltimore: Morgan Kaufmann*.

Gatsoulis, Y.; Alomari, M.; Burbridge, C.; Dondrup, C.; Duckworth, P.; Lightbody, P.; Hanheide, M.; Hawes, N.; Hogg, D.; Cohn, A.; et al. 2016. QSRlib: A Software Library for Online Acquisition of Qualitative Spatial Relations from Video. *In Workshop on Qualitative Reasoning (QR16), at IJCAI*.

Kalchbrenner, N.; Oord, A.; Simonyan, K.; Danihelka, I.; Vinyals, O.; Graves, A.; and Kavukcuoglu, K. 2017. Video Pixel Networks. In *International Conference on Machine Learning*.

Koppula, H. S.; Gupta, R.; and Saxena, A. 2013. Learning Human Activities and Object Affordances from RGB-D videos. *The International Journal of Robotics Research*, 32(8): 951–970.

LeCun, Y. A.; Bottou, L.; Orr, G. B.; and Müller, K.-R. 2012. Efficient Backprop. In *Neural networks: Tricks of the trade*, 9–48. Springer.

Lotter, W.; Kreiman, G.; and Cox, D. 2016. Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning. *arXiv preprint arXiv:1605.08104*.

Randell, D. A.; Cui, Z.; and Cohn, A. G. 1992. A Spatial Logic based on Regions and Connection. *KR*, 92: 165–176.

Sridhar, M.; Cohn, A. G.; and Hogg, D. C. 2010a. Relational Graph Mining for Learning Events from Video. In *Proceedings of the 2010 conference on STAIRS 2010: Proceedings of the Fifth Starting AI Researchers Symposium*, 315–327.

Sridhar, M.; Cohn, A. G.; and Hogg, D. C. 2010b. Unsupervised Learning of Event Classes from Video. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, 1631–1638. AAAI Press.

Srivastava, N.; Mansimov, E.; and Salakhudinov, R. 2015. Unsupervised Learning of Video Representations using LSTMs. In *International conference on machine learning*, 843–852. PMLR.

Van de Weghe, N.; Cohn, A.; De Tre, G.; and De Maeyer, P. 2006. A Qualitative Trajectory Calculus as a Basis for Representing Mmoving Objects in Geographical Information Systems. *Control and Cybernetics*, 35(1): 97–119.

Van Rijsbergen, C. 1979. Information retrieval: theory and practice. In *Proceedings of the Joint IBM/University of Newcastle upon Tyne Seminar on Data Base Systems*, volume 79.

Villegas, R.; Yang, J.; Hong, S.; Lin, X.; and Lee, H. 2017. Decomposing Motion and Content for Natural Video Ssequence Prediction. *arXiv preprint arXiv:1706.08033*.

Wang, Y.; Jiang, L.; Yang, M.-H.; Li, L.-J.; Long, M.; and Fei-Fei, L. 2018. Eidetic 3D LSTM: A Model for Video Prediction and Beyond. In *International Conference on Learning Representations*.

Xingjian, S.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; and Woo, W.-c. 2015. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In *Advances in Neural Information Processing Systems*.

Xu, J.; Ni, B.; Li, Z.; Cheng, S.; and Yang, X. 2018. Structure Preserving Video Prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.