Research Area: Deep Learning, Computer Vision

“Novel View Synthesis from Crowd Sourced Videos”

Since past few years, we have seen popularity of smartphones has increased dramatically. As a result lot of multimedia data (images/videos in particular) is generated from these devices. One can also record sensor data such as GPS location, Compass orientation of the smartphones while recording a video or capturing a photograph. This additional data can be used in a novel way for solving conventional problems. For example, we have implimented two algorithms for tracking the *event* using only the sensor data of GPS and compass in the smartphones, when the same event is video recorded by many cameras at the same time [1,2]. Here we define the event to be any activity, such as protest march, a live concert and so on, which is recorded by many people at the same time. Fig. 1 below shows a typical event recording scenario (an illustration) where five cameras are video recording a person moving around in a lawn. The underlying hypothesis here is that, at any given instant of time, there will be quite a few users pointing their cameras at the event. Hence one can use location and azimuth data from all the cameras to locate the event. Our result shows successful event localization using only sensor data.

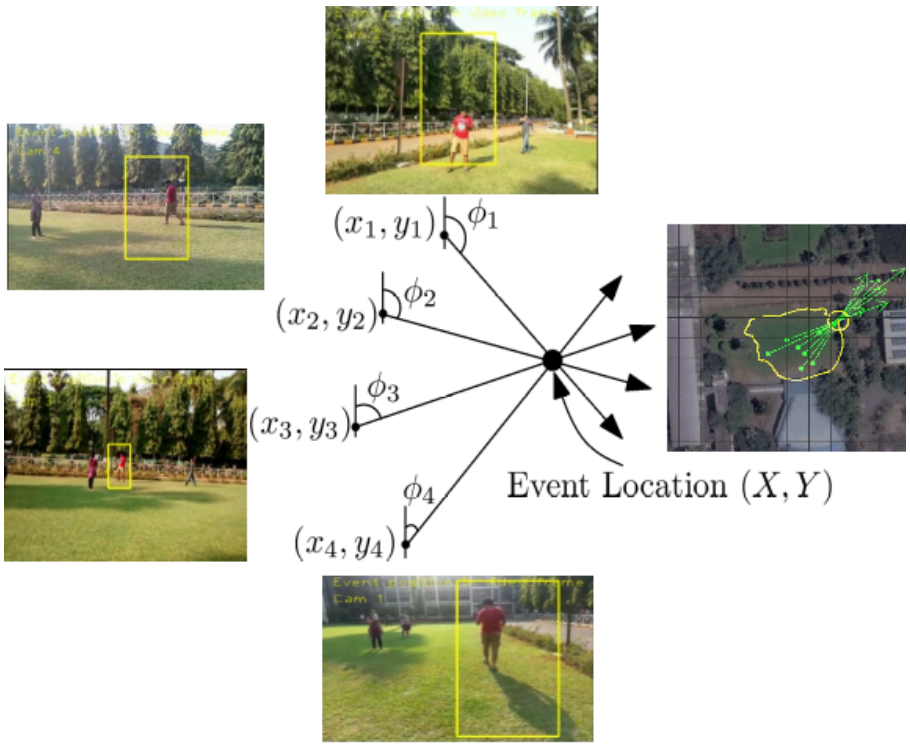

Figure 1: Typical event recording scneario

As the next part, currently we are working on synthesising novel view of the event which is been localized by previous algorithms. In particular, we know that event is being video recorded by many cameras at any given instant of time. And we have camera location and azimuth information available. Moreover, we also have event location avialable from the previous work. Hence, in theory, one can use this information to predict new, imaginary position for the camera from where event can be visualized in much better way. Our aim is to synthesize the event view from this new position using deep learning (DL). The primary motivation for this is the incredible success shown by DL in recent years for solving several challenging problems.

The overall setup of the proposed solution to the problem is as follows. We have multiple views (images/frames of videos) of the event and corrosponding camera location and orientation data. Given the new (imaginary) camera location and orientation one have to generate how the event will look like from this new camera viewpoint. We propose that given all the sensory and visual data, suitable DL architecture will be able to do so [3].

Currently, we have proposed an architecture which uses Spatial Transformer Networks (STN) [4, 5]. Its a DL architecture that will predict parameters of affine transformations to be applied on input images. These parameters are *data dependant* i.e, for each input a suitable set of parameters are predicted so that affine transformed image using these

parameters, will be better suited than the original image for given application/problem. Our hypothesis is that, we can train these STNs to generate several affine transformations per view and the novel view to be synthesized can then be computed using convex combination of all these transformed images. The proposed architecture have two separate pipelines closely interlinked with each other. One pipeline will predict paramters of the STN and the other will predict corresponding mask for convex combination. The overall architecture is shown in Fig. 2. The major challange for this task is the nature of visual data itself. Unlike previous works, we aim to solve this problem for a data collected from corw-sourced environment. As this data is collected from general crowd, each camera involved will have different intrinsic and extrinsic paramters, noisy sensor data, different viewing resolution, color saturation, etc. These settings make this problem more challenging and difficult than conventional problems of view synthesis addressed by research community in computer vision.

Further, as this approach sythesizes new view using available views, naturally, it can not synthesis parts of the scene which are occluded or poorly visible. Moreover, particular instanse of the problem we are attempting to solve falls under the catagory of wide-baseline camera setup which is known to be further challanging. This motivated us for using Generative Adversarial Networks (GAN) for view synthesis which will be the later part of the work.

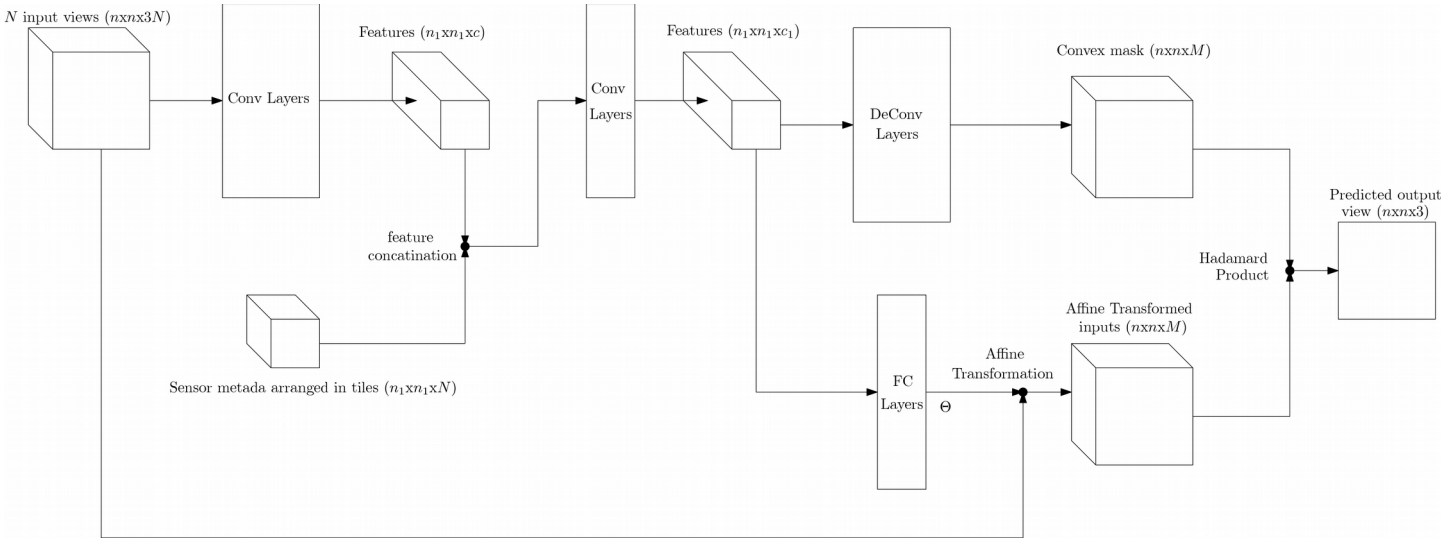

Figure 2: Proposed architecture for novel view generation

References:

[1] More, Amit, and Subhasis Chaudhuri. "Event geo-localization and tracking from crowd-sourced video metadata." *Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing*. ACM, 2016.

[2] More, Amit, and Subhasis Chaudhuri. "Pseudo-likelihood Approach For Localizing The Events using Sensor-Metadata From Crowd-sourced Videos." *Under review.*

[3] Flynn, John, et al. "DeepStereo: Learning to predict new views from the world's imagery." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

[4] Jaderberg, Max, Karen Simonyan, and Andrew Zisserman. "Spatial transformer networks." *Advances in Neural Information Processing Systems*. 2015.

[5] Finn, Chelsea, Ian Goodfellow, and Sergey Levine. "Unsupervised learning for physical interaction through video prediction." *Advances in Neural Information Processing Systems*. 2016.