Deep Learning for Augmented Reality

Jean-François Lalonde
Université Laval
Quebec City, Canada
jflalonde@gel.ulaval.ca

***Abstract*—Augmented reality aims to mix real-world visual content with virtual objects. Achieving realistic results involves solving challenging computer vision tasks, such as tracking real 3D objects and estimating the illumination conditions of a scene. In this short paper, we present how these two challenging tasks can be solved robustly and accurately with deep learning. In both cases, deep convolutional neural networks are trained on large amounts of data, and achieve state-of-the-art results.**

***Index Terms*—deep learning, augmented reality, tracking, lighting estimation**

I. Introduction

Augmented reality (AR) aims to mix real-world visual content (photos, movies, etc.) with virtual objects. While doing so has so far been confined to the realm of visual effects artists, AR is now on the verge of becoming part of our everyday lives. Indeed, due in most part to recent progress in SLAM-based camera localization techniques which can robustly position a camera in an unknown 3D environment, devices supporting various forms of AR are now commercially available and are paving the way to a wide range of applications. Unfortunately, the results obtained from these devices are a long distance away from matching the high degree of realism attained by special effects artists, who routinely fool audiences into thinking that computer-generated objects are real.

The key to achieving realism when compositing (mixing) virtual with real content is that the virtual object must *share the same characteristics* as the real world. Let us illustrate this idea with the example of placing a virtual apple onto a real plate. First, the apple must be stay on the plate irrespective of the user point of view. Second, if the user moves the plate, then the apple must respond to that movement accordingly (by following the plate, or, depending on the user skills, by falling from it!). Third, the apple must be lit in the same way as its real surroundings to realistically blend in.

All three problems involve solving three challenging computer vision problems: 1) camera localization, 2) object tracking, and 3) illumination estimation. While robust solutions exist for the first, the other two are still very much open research problems. In this short paper, we will briefly present how these two challenging tasks can be solved robustly and accurately with the use of deep learning [1].

II. Learning to track objects in 3D

We present an accurate, real-time temporal 6-degrees of freedom (DOF) object tracking method which is more robust to occlusions than existing state-of-the-art algorithms. Our main key contribution is to frame 6-DOF tracking as a deep learning problem. This contribution provides us with three key benefits. First, deep learning architectures can be trained on very large amounts of data, so they can be robust to a wide variety of capture conditions such as color shifts, illumination changes, motion blur, and occlusions. Second, they possess very efficient GPU implementations that can be processed in real-time on mobile GPUs given a small enough network. Finally, and perhaps most importantly, no hand-designed features need to be computed: object-specific features can automatically be learned from data. This is in contrast to most previous work (e.g. [2], [3]) which compute specific, hand-designed features.

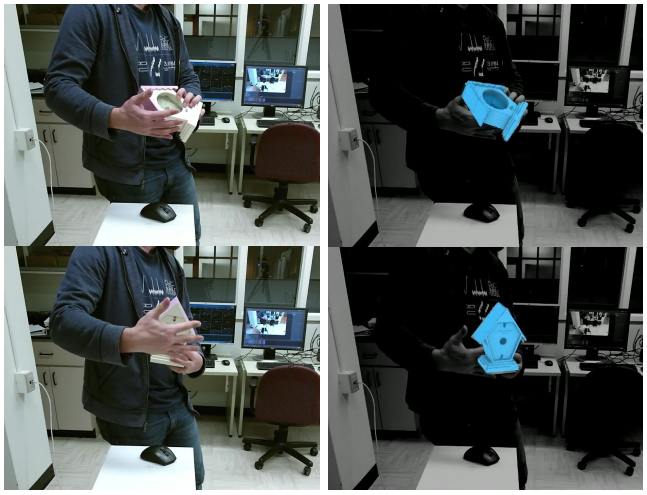

Fig. 1. Tracking objects in real-time with deep learning. Left: input image, right: 3D model overlaid on the object.

Applying a deep convolutional neural network (CNN) to tracking is not trivial. Indeed, temporal tracking differs from tracking by detection in that the temporal tracker uses two frames (images) adjacent in time, and assumes knowledge of the object pose at the previous frame. To train a deep network on that task, one could straightforwardly use the current and previous frames directly as input. Unfortunately, while doing so yields low prediction errors on a "conventional" machine learning test set (composed of pairs of frames as input and rigid pose change as target), it completely fails to track in sequences of multiple frames. Indeed, since the network never learned to correct itself, small errors accumulate

The author gratefully acknowledges the financial support of NSERC, FRQ-NT, Adobe, Creaform, and Nvidia that supported this research.

Fig. 2. Our method learns a direct mapping from image appearance to scene lighting from large amounts of real image data; it does not require any additional scene information, and can even recover light sources that are not visible in the photograph, as shown in these examples. Using our lighting estimates, virtual objects can be realistically relit and composited into photographs.

rapidly and tracking is lost in a matter of milliseconds. Another solution could be to provide the previous estimate of the pose change instead of the previous frame as input to the network. In this case, this information alone is not rich enough to enable the network to learn robust high level representations, also yielding high tracking errors. To solve this problem, we propose to use an estimate of the object pose from the previous timestep in the sequence as input to the network, in addition to the current frame. This allows the network to correct errors made in closed loop tracking. The feedback, which is the estimate of the current object pose, is obtained by rendering a synthetic frame of the tracked object. Thus, our approach requires a 3D model of the object a priori, and the tracker is trained for a specific object. To the best of our knowledge, we are the first to use deep learning for 6-DOF temporal object tracking.

In a nutshell, our deep neural network accepts two inputs: an image of the object rendered at its predicted position (from the previous timestamp in the video sequence), and an image of the observed object at the current timestamp. The network directly outputs the 6 degrees of freedom (3 for translation, 3 for rotation in Euler angles) representing the pose change between the two inputs. To train the network, we rely on a dataset of synthetically-generated images of the object, obtained from its 3D model. When evaluated on a large dataset of real objects, we find our approach is more stable (0.4 mm movement per frame in a static video vs. 1.2 mm for [2]), more robust under significant occlusions (12.5 mm average error at 45% occlusion vs. 138 mm for [2]), and more accurate at large object speed (3.6° average error vs. 8.1° for [2]), see tab. I for more details. Fig. 1 shows a qualitative example of our tracker in action. More details about this work can be found in [4], [5].

Method	Stability		45% occlusion		Fast speed	
	t (mm)	r (°)	t (mm)	r (°)	t (mm)	r (°)
Ours	0.56	0.52	12.5	10.0	11.1	3.6
[2]	1.20	1.30	138	70.3	10.7	8.1

TABLE I
QUANTITATIVE PERFORMANCE OF OUR TRACKER, COMPARED TO PREVIOUS WORK [2]. OUR METHOD ACHIEVES SUPERIOR TRANSLATION (T) AND ROTATION (R) MEAN ERROR, IN TERMS OF STABILITY, OCCLUSION, AND FAST OBJECT MOTION.

III. LEARNING TO ESTIMATE LIGHTING

Inferring scene illumination from a single photograph is a challenging problem. The pixel intensities observed in an image are a complex function of scene geometry, materials properties, illumination, the imaging device, and subsequent post-processing. Disentangling one of these factors from another is an ill-posed inverse problem. This is especially hard from a *single limited field-of-view image*, since many of the factors that contribute to the scene illumination are not even directly observed in the photo (fig. 2). This problem is typically addressed in two ways: first, by assuming that scene geometry (and/or reflectance properties) is given (either measured using depth sensors, reconstructed using other methods, or annotated by a user), and second, by imposing strong low-dimensional models on the lighting.

In this work, we propose a method to infer high dynamic range (HDR) illumination from a single, limited field-of-view, low dynamic range (LDR) photograph of an indoor scene. Our goal is to be able to model the range of typical indoor light sources, and choose a spherical environment map representation that is often used to represent real-world illumination [6]. We also want to make this inference robust to errors in geometry, surface reflectance, and scene appearance models. To this end, we introduce an end-to-end learning based approach, that takes images as input and predicts illumination using a deep neural network.

We use a convolutional neural network that takes the photo as input, produces a low-dimensional encoding of the input through a series of convolutions downstream and splits into two upstream expansions, with two distinct tasks: (1) light intensity estimation, and (2) RGB panorama prediction. To train the network, we rely on a large dataset of panoramas [7]. Given a panorama, we extract a regular image assuming a perspective camera model with randomly-sampled parameters, and use it as input to the neural network.

We evaluate our technique through a user study, in which participants were asked to choose which of two images was the most realistic: the image containing a virtual object relit with 1) the ground truth illumination, and 2) the estimated illumination conditions. The results indicate that renderings obtained with our estimated illumination were considered as

or more realistic than the ground truth result in 41.8% of the responses, which is a significant improvement over the previous work that reached, at most, a performance of 27.7%. More details about this work, as well as variants that are adapted to outdoor lighting, can be found in [8]–[10].

IV. Discussion

In this paper, we presented techniques which rely on deep learning to solve two challenging computer vision problems, namely 6-DOF object tracking and illumination estimation. In both cases, deep convolutional neural networks are trained on large amounts of data, and achieve state-of-the-art results. We believe that such solutions should enable much more realistic AR applications, which will be able to adapt to rapid scene changes and to challenging illumination conditions.

Acknowledgements

The author wishes to thank his students (Mathieu Garon, Marc-André Gardner, Yannick Hold-Geoffroy, Jinsong Zhang) and collaborators (Denis Laurendeau, Christian Gagné, Kalyan Sunkavalli, Sunil Hadap, Emiliano Gambaretto, Ersin Yumer, and Xiaohui Shen) who have all contributed to this work.

References

[1] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1.

[2] D. J. Tan, F. Tombari, S. Ilic, and N. Navab, "A versatile learning-based 3D temporal tracker: Scalable, robust, online," in *IEEE International Conference on Computer Vision*, 2015.

[3] D. J. Tan, N. Navab, and F. Tombari, "Looking beyond the simple scenarios: Combining learners and optimizers in 3D temporal tracking," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 11, pp. 2399–2409, 2017.

[4] M. Garon and J.-F. Lalonde, "Deep 6-DOF tracking," *IEEE Transactions on Computer Graphics and Visualization*, vol. 23, no. 11, 2017.

[5] M. Garon, D. Laurendeau, and J.-F. Lalonde, "A framework for evaluating 6-DOF object trackers," in *European Conference on Computer Vision (ECCV)*, 2018.

[6] P. Debevec, "Rendering synthetic objects into real scenes : Bridging traditional and image-based graphics with global illumination and high dynamic range photography," in *Proceedings of ACM SIGGRAPH*, 1998.

[7] J. Xiao, K. A. Ehinger, A. Oliva, and A. Torralba, "Recognizing scene viewpoint using panoramic place representation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[8] M.-A. Gardner, K. Sunkavalli, E. Yumer, X. Shen, E. Gambaretto, C. Gagné, and J.-F. Lalonde, "Learning to predict indoor illumination from a single image," *ACM Transactions on Graphics (SIGGRAPH Asia)*, vol. 9, no. 4, 2017.

[9] Y. Hold-Geoffroy, K. Sunkavalli, S. Hadap, E. Gambaretto, and J.-F. Lalonde, "Deep outdoor illumination estimation," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[10] J. Zhang and J.-F. Lalonde, "Learning high dynamic range from outdoor panoramas," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.