Research Statment

Xue, Tianfan

My research focuses on visual correspondence between images. First, **visual correspondence between images reveals various physical properties of our visual world**. For example, correspondences between two temporal adjacent frames reveal how objects move over time, and many motion estimation algorithms [1, 2] are based on temporal correspondence. Correspondences between images captured from different viewpoints reveal the 3D locations and geometries of objects, as well the camera motion between different viewpoints. Most of 3D estimation algorithms, like structure from motion [4] and dense stereo matching are based on spatial correspondence [3].

Second, **inferred properties from the visual correspondence can be used to manipulate captured images or videos**. For example, to rotate an object, like a chair or a cup, in a captured image or a video, we need to know the 3D geometry of that object. To edit or synthesize a realistic movement of objects in a video, we need to first extract motion field of similar objects in a reference video.

My research goal is, through the correspondence between images, to understand underlying structures of our visual world, and to assist the editing of captured images and videos.

Research Contributions

Improved Visual Correspondence Estimation Traditional motion estimation [1, 2] and 3D reconstruction algorithms [4] are based on visual correspondence between images, and there are still two main challenges in finding correspondence between images. First, the state-of-the-art flow and stereo algorithms can sometimes introduce errors in less textured regions. Second, most of matching algorithms are based on the brightness constancy assumption, that it the intensity of two matched pixels should have the same or similar intensity. However, such assumption only holds for solid objects, and it cannot be applied to semi-translucent objects, like fluid flow.

Figure 1: Edge-based stereo matching. We match intensity edges in a multi-frame sequence (a) to obtain a sparse depth map (b), from which we infer the dense depth map (c).

First, to improve the quality of stereo matching, we proposed an edge-based stereo matching algorithm [6] to deal with less textured regions and the "foreground-fattening effect" (Figure 1). The basic idea is that in the first round of the algorithm, we only consider pixels on edges, either on texture edges or occlusion edges. It allows us to focus the computation on the important features, to identify object boundaries early, and to defer reasoning in untextured areas. Once we find robust matches of those pixels on edges, and we then fit overlapping local planes to these coarse matches. The final depth map is recovered by assigning each pixel to one of detected planes. Results show this edge-based approach can create a clean depth map with sharp boundaries between objects.

Second, to find robust matches for refractive fluid objects, we design a new flow algorithm that extends the traditional brightness constancy assumption. Although the fluid objects themselves can be hardly tracked visually, its motion causes small intensity variations of the background. One main observation is that such intensity variations are consistent over small space-time volumes (Figure 2). We call these intensity variations refraction wiggles, and use them as features for tracking and stereo fusion to recover the fluid motion and depth from video sequences. We designed algorithms both for 1) measuring the (2D, projected) motion of refractive fluids in monocular videos, and 2) recovering the 3D position of fluid objects from stereo cameras [8].

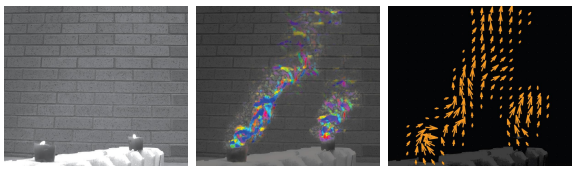

Figure 2: Measuring the velocity of imperceptible candle plumes from standard videos. The heat rising from two burning candles (a) cause small distortions of the background due to refraction. We proposed an algorithm can both visualize the air flow and measure its 2D motion (c).

Novel Visual Correspondence Applications Once we infer the correct correspondence between images, many image and video editing problems become much easier

and approachable. For example, from just a single view, it is hard to get a clean segmentation of foreground objects from background, and possibly change or remove foreground objects. However, if we take a short video sequence by moving the camera, objects at different layers will match pixels in different locations in other frames, and it is much easier to separate them.

One application is to remove the visual obstruction from captured videos [7]. For example, when taking pictures through glass windows, reflections from indoor objects can obstruct the outdoor scene we wish to capture, as shown in Figure 3. To remove these visual obstructions, we instruct the user to take a short image sequence while slightly moving the camera. Our key observation is that the reflecting or obstructing planes usually have different depth from that of the main scene, and thus different motion pattern in the captured sequence. Thus, we can separate the main scene from the obstruction based on motion parallax, and fill the holes left by the obstruction layers by aggregating information from corresponded pixels in other frames. In this way, we can deal with obstruction appeared at various scenarios, including shooting through reflections, fences, and raindrop-covered windows.

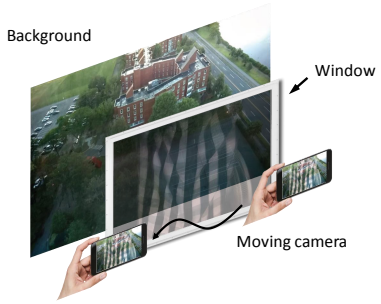

Figure 3: Remove visual obstructions from videos captured by a moving camera.

Furthermore, given visual correspondences extracted from a large number of videos, we can learn how to synthesize new videos from a single input image. In contrast to traditional methods that are either deterministic or non-parametric, we propose to model future frames in a probabilistic manner [9], so shat we can synthesize many possible future movement of an object. We propose a novel image synthesis network, which first chops the input image into different segments and synthesize a new image by moving segments and combining them together. Our algorithm can automatically learn the correspondence between two temporally neighboring frames in the training videos. At testing time, given an input image, it can also sample the correct movement of each segment, based on learned correlation between images and motion fields in the training set.

Figure 4: Predicting the movement of an object from a single snapshot (a) is often ambiguous. Therefore, we proposed a probabilistic, content-aware motion prediction model that learns the conditional distribution of future frames

Future Research Plan

Holistic visual world understanding through correspondence So far, most work separately infer motion or 3D geometry of objects from either temporal or spatial correspondence between images. Actually, these two problems are intertwined. For example, when we capture a video of a static scene, the observed motion on the image plane depends on the depth of the scene, so that we can infer the 3D geometry from camera motion. Reversely, when we stereo-match a set of images from a dynamic scene taken at different timestamps, we also need to infer how much objects move between frames. Therefore, a principle way to solve the problem is to infer both 3D geometry and movement of objects simultaneously, and researchers also start to push this direction [5].

Furthermore, besides motion and 3D, we can also infer other properties of our visual world through correspondence. For example, from a set of images under different lighting conditions, we can also infer the light sources, as well as the bidirectional reflectance distribution function (BRDF) of each object in the scene. Also, from videos of moving fluid, we can also infer some physical properties of the fluid, like viscosity or transmittance.

Clip-based image and video editing and synthesis All these examples shown above illustrate that the information we can get from a short video clip is much richer than that from a single image, including rough 3D geometry of the scene, dynamics of objects, lighting conditions, etc. These rich information about our visual world can greatly improve the performance of existing image and video editting algorithms, and make many other applications possible. For example, we can easily manipulate objects in captured sequences, once we know their

3D geometry. With estimated 3D geometry and motion from a short video clip, we can also synthesize a stereo sequence, or a full light-field sequence, that can be viewed interactively either on mobile devices, or VR headsets.

References

[1] Horn, B.K., Schunck, B.G.: Determining optical flow. Artificial intelligence 17(1-3), 185–203 (1981)

[2] Lucas, B.D., Kanade, T., et al.: An iterative image registration technique with an application to stereo vision. In: IJCAI. vol. 81, pp. 674–679 (1981)

[3] Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International journal of computer vision 47(1-3), 7–42 (2002)

[4] Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3D. In: ACM SIGGRAPH. pp. 835–846 (2006)

[5] Vogel, C., Schindler, K., Roth, S.: Piecewise rigid scene flow. IJCV pp. 1377–1384 (2013)

[6] Xue, T., Owen, A., Geosele, M., Scharstein, D., Szeliski, R.: Multi-frame stereo matching with edges, planes, and superpixels. in preparation (2017)

[7] Xue, T., Rubinstein, M., Liu, C., Freeman, W.T.: A computational approach for obstruction-free photography. ACM SIGGRAPH 34(4) (2015)

[8] Xue, T., Rubinstein, M., Wadhwa, N., Levin, A., Durand, F., Freeman, W.T.: Refraction wiggles for measuring fluid depth and velocity from video. ECCV (2014)

[9] Xue, T., Wu, J., Bouman, K., Freeman, B.: Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In: NIPS. pp. 91–99 (2016)