Scale-invariant Object Recognition from Industrial Light-field imaging

S. Cloix, D. Hasler, T. Pun •

A new approach to scale-invariant object recognition based on bag-of-visual-words is presented. Our method was tested on our new database of light-field images and assessed an excellent recognition rate greater than 90% despite a scale variation of about 200%. Our versatile light-field image dataset, CSEM-25, composed of 5 classes captured under several poses and backgrounds, was built with an industrial plenoptic camera, the Raytrix R5. It will be made available for research purposes.

Light-field imaging is an enhancement of conventional imaging that records not only the intensity of light but also the direction of every light ray hitting a camera [1]. In practice, the recording of a light field is either performed using arrays of cameras, or – as in our case – using a dedicated camera that includes a microlens array, also called a *plenoptic camera*. In our tests we used the industrial 4-megapixel Raytrix [2] camera.

The existing light field datasets, mostly capturing one object, are seldom suitable for classification purposes but 3D reconstruction. We aim at providing a dataset exhaustive enough to be used for many different vision and classification tasks. The dataset is composed of 5 classes of 5 objects of known and similar size. For each object, pose (72 angles) and distance (21 from 28 cm to 50 cm), four captures are acquired: two with a uniform background and two with a landscape background randomly picked from a database of high resolution images (Figure 1).

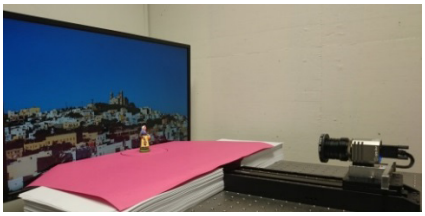

Figure 1: Our acquisition setup composed of a motorized linear stage, a motorized turntable, a background screen and a uniform colored ground.

The image recorded by the Raytrix camera is a group of micro-images lying on a hexagonal grid. An interesting feature of the recorded image is that the content of each micro-image does not vary a lot when highly increasing the distance of the object from the camera; only the number of times a pattern appearing inside a micro-image varies significantly (Figure 2) across neighbouring micro-images. We therefore aim at taking advantage of these pattern repetitions and the small variation in scale within the micro-images to develop a recognition system that is invariant to the scale induced by the distance.

Our object recognition method is based on bag-of-visual-words strategy: (i) a codebook is built from an unsupervised clustering method and (ii) is used to build a histogram of each image. The histograms of the test images are then compared to each of the training images of the labelled objects.

The codebook is a set of whitened pixel patches learnt from small patches extracted within each micro-image of a training-image set. The training set is made of segmented captures of each object at the closest distance. A histogram is then extracted for each object. For each test image, the small patches are extracted at a fixed location within a fixed region of interest. Each bin of a histogram represents the number of occurrence of the corresponding visual word in the region of interest. The farther the object, the smaller the number of visual words belonging to the object. As the histogram of a test image is expected to have the same shape than of the training image but with lower amplitude, we scale up the test histogram and compare it with the histogram of the training images by minimizing a thresholded ℓ_1 distance.

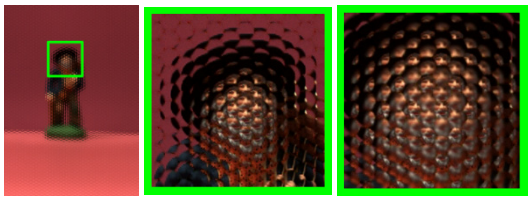

Figure 2: Dataset samples: on unified background, an instance of the "person" class is captured from the farthest distance (left), a zoom region (middle) of the left image and a zoom region at a closer distance to the camera.

We evaluate our approach on five objects, one instance of each class. The training images are composed of the ones with a uniform background. The test images are from four subsets, the first two with a uniform background and the last two with random backgrounds. At the closest distance, we obtain a recognition rate of 100%, the background not having a large impact. Using a fixed size detection window, the farther the objects, the lower the recognition rate, due to the noise introduced by the background that fills an increasing proportion of the detection window. We exceed 90% of correct recognition for each tested distance, the recognition rate expectedly decreasing with the distance (from 100% at the closest distance to 90% at the farthest one).

From the properties of our industrial plenoptic camera we designed a new real-time recognition approach that is robust to large scale variation of almost twice the size of the object when farthest from the camera. With a codebook of a few words (100 visual words), we reached a recognition rate greater than 90%. As next steps, we aim at scaling up the system to recognize more objects and also to classify objects by category. The dataset is available on demand.

This work is co-funded by the Swiss Hasler Foundation SmartWorld Program, grant Nr. 11083.

• Computer Science Department, CVML, University of Geneva, Switzerland

[1] E. H. Adelson, J. R. Bergen, "The plenoptic function and the elements of early vision", Computational models of visual processing, 1 (2) (1991) 2

[2] www.raytrix.de