Real-time Scale-invariant Object Recognition from Light Field Imaging

Cloix, Séverine; Pun, Thierry; Hasler, David

How to cite

CLOIX, Séverine, PUN, Thierry, HASLER, David. Real-time Scale-invariant Object Recognition from Light Field Imaging. In: Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISAPP 2016. Rome (Italy). [s.l.] : SCITEPRESS - Science and Technology Publications, 2016. p. 336–344. doi: 10.5220/0005678603360344

This publication URL: https://archive-ouverte.unige.ch//unige:145451
Publication DOI: 10.5220/0005678603360344

Real-time Scale-invariant Object Recognition from Light Field Imaging

Séverine Cloix[1,2], Thierry Pun[2] and David Hasler[1]
[1]*Vision Embedded Systems, CSEM SA, Jaquet Droz 1, Neuchâtel, Switzerland*
[2]*Computer Science Department, University of Geneva, Route de Drize 7, Carouge, Switzerland*

Keywords: Object Recognition, Object Classification, Light Field, Plenoptic Function, Scale Invariance, Real-time, Dataset.

Abstract: We present a novel light field dataset along with a real-time and scale-invariant object recognition system. Our method is based on bag-of-visual-words and codebook approaches. Its evaluation was carried out on a subset of our dataset of unconventional images. We show that the low variance in scale inferred from the specificities of a plenoptic camera allows high recognition performance. With one training image per object to recognise, recognition rates greater than 90 % are demonstrated despite a scale variation of up to 178 %. Our versatile light-field image dataset, CSEM-25, is composed of five classes of five instances captured with the recent industrial Raytrix R5 camera at different distances with several poses and backgrounds. We make it available for research purposes.

1 INTRODUCTION

The detection of "everyday objects" is still an active area of research. Current algorithms that detect objects by category, for example keys, doors, cars, cats, glasses, etc. are still not good enough for practical applications (Everingham et al., 2015) that require very low false positive even nil according to the application. While most of the works on classification has been dealing with 2D images over the last decades, a number of public datasets has been made available for the purpose of developing new methods. A single 2D image however gives partial information of the fronting scene since the sensor records a projection of the 3D scene, losing the third spatial dimension.

The changes in distance of an object of known real dimensions are represented by a change in size on the 2D image. The scale is thus often dealt with iteratively running the detector on downsampled images or with depth estimation (Gavrila and Munder, 2006) prior to the recognition step (Helmer and Lowe, 2010). Stereo-view strategies allow to extract the third spatial dimension, resulting in either a sparse 3D point cloud (Cloix et al., 2014) or a range of distances to infer the size of the detection window in object detection algorithms (Helmer and Lowe, 2010). The depth map of the captured scene is computed by triangulating over the scene points visible and identified in both images. Commercial devices commonly employed are stereo cameras like the Point Grey bumblebee2[1] and active sensors, e.g. Microsoft Kinect[2] and Asus Xtion[3], developed to cope with textureless scenes.

With more than two views, we can call the whole capture a subset of the "light field". The definition of light field comes from the plenoptic function (Adelson and Bergen, 1991). For each point in the 3D scene, the intensity distribution is

$$P(\theta, \phi, \lambda, t, V_x, V_y, V_z), \tag{1}$$

where θ and ϕ are the spherical coordinates of the direction of the light ray, λ the wavelength and t the time dimension. The viewpoint is defined by V_x, V_y and V_z. In practice, a conventional camera is capable of recording a 2D slice of the scene irradiance. With the multi-view strategy, we are able to add three other dimensions describing the location of the view point. Light fields are thus captured with an array of cameras, a gantry (Levoy, 2011) or the use of a turntable and a robot (Zobel et al., 2002). Another way to augment the conventional image capturing with two dimensions is the use of a microlens array. This 4D parameterization (x, y, u, v) is done by two planes, the

[1]http://www.ptgrey.com/bumblebee2-firewire-stereo-vision-camera-systems

[2]http://www.xbox.com/en-US/xbox-one/accessories/kinect-for-xbox-one

[3]http://www.asus.com/Multimedia/Xtion/

Cloix, S., Pun, T. and Hasler, D.
Real-time Scale-invariant Object Recognition from Light Field Imaging.
DOI: 10.5220/0005678603360344
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 4: VISAPP*, pages 336-344
ISBN: 978-989-758-175-5

viewpoints plane, (u,v), and the sensor plane, (x,y), and allows the measurement of the directional distribution of the light (Ng, 2005) (Figure 1). The latter is named the 4D plenoptic camera with commercial versions like Lytro[4] and Raytrix[5].

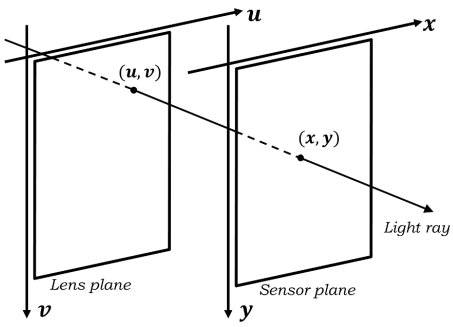

Figure 1: Two-plane parameterization of the 4D Light field as explained in (Ng, 2005).

In this article, we present a novel light field dataset as well as a scale-invariant object recognition system based on bag-of-visual-words and codebook approaches. We assess our method on a subset of our novel light field dataset. To the best of our knowledge, our dataset is the first of its kind in the domain of light field for computer vision. We believe that our approach is the first work that shows how industrial light field imaging can be successfully employed in object recognition. We thus expect it to become a baseline in the community.

This paper is organized as follows. Section 2 describes the state-of-the-art in light field vision related to object classification. A review on the existing light field image datasets is recalled in Section 3 followed by an exhaustive description of our proposed dataset (section 4), along with several intended usage scenarios of the data. Section 5 explains the approach of our scale-invariant recognition system from raw light field images captured by an industrial plenoptic camera. The experimental results are detailed and discussed in Section 6 before concluding on the future work in Section 7.

2 RELATED WORK

Object recognition and classification is an old topic in computer vision on which research keeps on progressing thanks to new algorithms and new sensors. Light field imaging is mainly employed for depth estimation and 3D scene or object reconstruction, although, as of today, it is still unclear in which domain the light field technology will stand out. The existing datasets were built according to the applications they were dedicated to. The first light field datasets were built by the Stanford Computer Graphics Laboratory and are used for computer graphics research [6]. Their array of cameras (Wilburn et al., 2005) captures scenes for 3D reconstruction like (Levoy, 2011). Disney research is also a big player in the domain of scene reconstruction (Kim et al., 2013). These datasets are however not used for classification purposes. To our knowledge the authors of (Ghasemi and Vetterli, 2014) are the first to exploit the light field images for object recognition. They present a scale-invariant feature, called STILT, built from Epipolar Plane Images (EPIs) to recognize buildings. An EPI, initially introduced in (Bolles et al., 1987), is the representation of the subset of the data in the $x-u$ domain, where each point represents a line whose slope is proportional to the depth. The STILT feature is drawn on the Hough transform line detection method and represents the signature of the entire light field captured, the invariance in scale being transposed in a multiplicative scalar. In (Ghasemi et al., 2014), this feature is used for object category classification. As far as classification methods are concerned, recent work of (Coates et al., 2012) shows state-of-the-art results in unsupervised feature learning on conventional 2D images. One of the learning phase is based on k-means dictionary learning, similar to codebook learning from bag-of-visual-word models (Csurka et al., 2004).

Looking at captures from the Raytrix camera, one striking characteristic in the light field images of an object at two different distances is the redundancy of a unique point of the scene on the final image: the closer the object, the greater the redundancy. We aimed at taking advantage of this redundancy feature to build a scale-invariant object recognition system.

In this paper we exploit the redundancy characteristic of the Raytrix camera to build a recognition system of objects at various distances without explicitly estimating the distance as a feature, nor downsampling the input capture. Another contribution is the release of a new light field images dataset for object class recognition purposes as described in 4.4.

3 EXISTING LIGHT FIELD DATASETS

In order to reduce the assumptions related to physical

[4]https://www.lytro.com/
[5]http://www.raytrix.de/
[6]http://lightfield.stanford.edu/

hardware and to keep focus on the development of new algorithms, other datasets were made of synthetic images using Blender™[7] (Wetzstein,; Wanner et al., 2013). Authors of (Wanner et al., 2013) also created datasets capturing real scenes. Image matting is also an application for which (Joshi et al., 2006) and (Cho et al., 2014) created their own datasets from real scenes.

As far as object classification is concerned, it is still a challenging research topic and several datasets dedicated to competitions were built. (Ghasemi et al., 2014) reviewed the most employed ones. The use of light field imaging for object class recognition is however rather new and to our knowledge, LCAV-31 is the only dataset that shares several specifications with ours. Dedicated to object class recognition, LCAV-31 gathers 31 classes of household and office objects captured with a Lytro camera. Each instance is captured with 3 different viewpoints and several object locations and angles. Each capture is converted from the Lytro format to a JPG image and is a 3010×3030 pixel grid of 10×10 sub-views. While our dataset detailed in Section 4 is based on a lens-grid representation, the LCAV-31 captures are view-grid-based ones, i.e. each light field is represented by an array of pictures. LCAV-31 does not offer metadatas such as location and pose nor masks for segmentation purposes.

4 PROPOSED DATASET

4.1 Goals and Specifications

We aim at building a multipurpose dataset of object classes to address several aspects of computer vision applications using light field with a lens-grid-based representation, namely:

- object classification,
- object recognition,
- corner detection,
- feature point extraction and tracking,
- (3D) pose estimation,
- 3D reconstruction.

4.2 The Acquisition Setup

We built an automated set-up composed of a motorized linear stage, a motorized rotary stage, a high resolution background screen and a Raytrix R5 camera.

[7] http://www.blender.org/

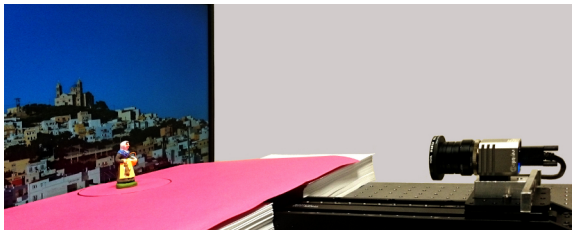

Figure 2: Acquisition setup. It is composed of a motorized linear stage, a motorized turntable, a high resolution background screen and a uniform colored ground. Here one instance of the "person" category is being captured with a randomly chosen background.

The ground is covered by a uniform colored paper. Each object is located in the middle of the turntable. It is captured from 0 to 355 degree with a 5 degree pitch and at 21 distances from the camera (from 28 to 50 cm). The turntable has an accuracy of 0.2 millidegrees, and the linear stage an accuracy of 2 microns. Thus the relative object pose from one image to any other in the dataset is known with high precision.

The object is placed in front of a background screen, which is located 6 cm behind the object center (Figure 2). For each object and each pose, four captures are acquired: two captures with a uniform background and two captures with a landscape background randomly picked from a database of 380 high resolution images. The screen resolution is large enough to avoid blur in the acquired image. The captures on uniform backgrounds are for computing a mask of the presence of the object, which can be used to place the object into a virtual environment. This operation of extracting the light field of the object and integrating it into the light field of a virtual scene being a complicated operation, we added the two captures with a high resolution background to make the database ready for object classification or detection on cluttered background. The dataset is composed of 5 classes of 5 instances of known size (less than 7 cm width, and height) in order to be in the field of view of the camera for all the distances (Figure 3). The categories are : person, four-legged animal, fruit, box and car.

The ground truth of our dataset, named CSEM-25, gives: (i) the category whose name belongs to the WordNet® lexical database, (ii) the angle, (iii) the distance from the camera and (iv) the intra-class instance number, (v) the number that refers to a miscellaneous background.

In order to enlarge the usage of our dataset, each object and each pose acquisition comes with (i) the raw colored images in PNG format and of size 2044×2044 pixels, (ii) its mask on raw data and (iii) the best "all-in-focus" image generated with the Raytrix SDK.

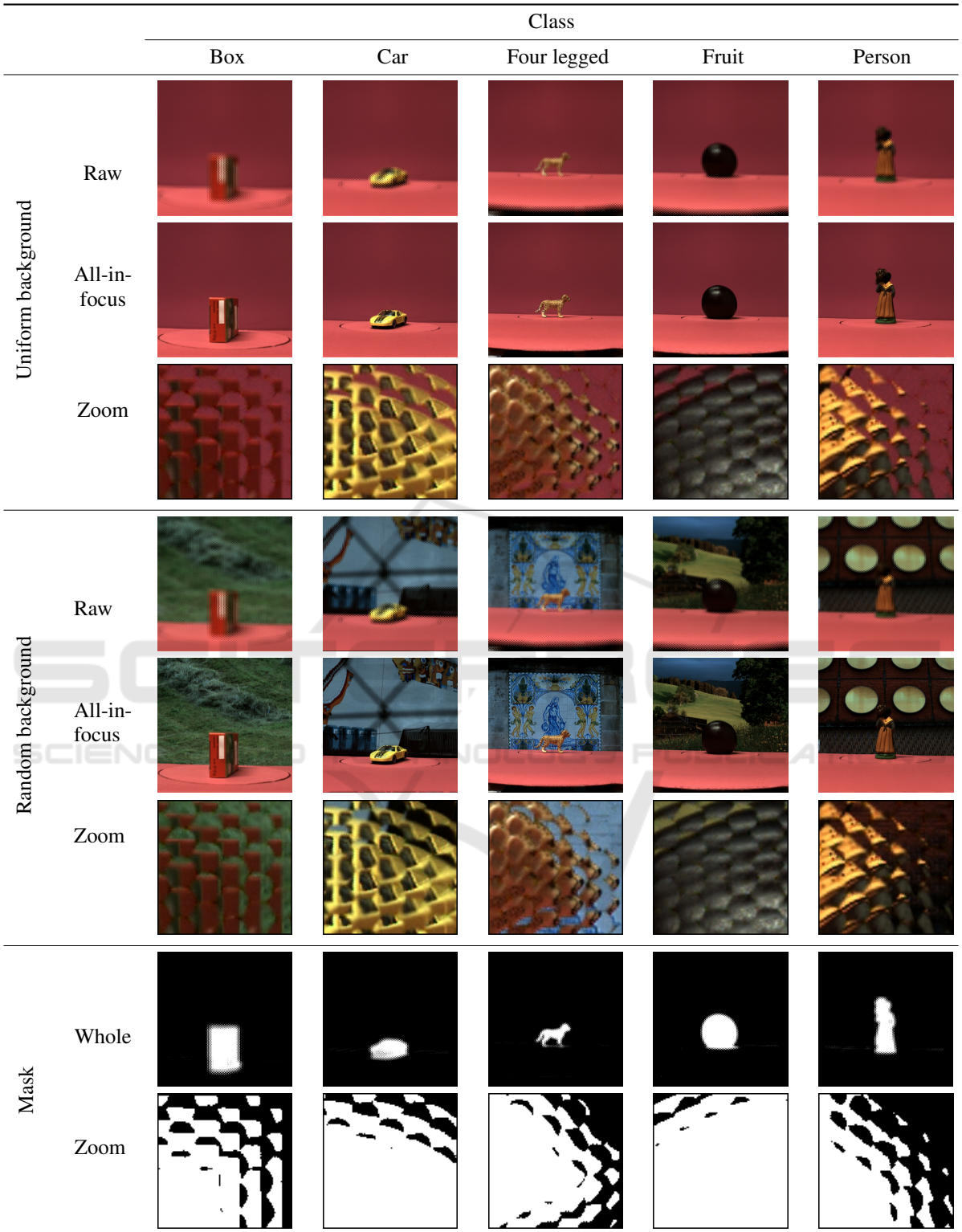

Figure 3: Dataset samples: on one of the uniform backgrounds and one of the random backgrounds, here are presented one instance of each class in one of the poses. The "*Raw*" rows show the colored raw image produced by the Raytrix camera. The "*All-in-focus*" rows show the processed image generated by the Raytrix SDK where each pixel is in focus. For each instance and pose, we deliver a mask as shown in row "*Mask*". The "*Zoom*" rows is a zoom in portion of the raw image or the mask.

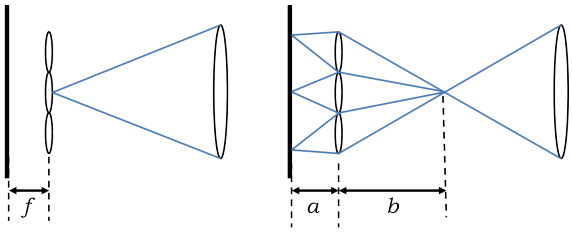

Figure 4: The difference between a Plenoptic 1.0 camera (left) and a Plenoptic 2.0 camera (right) lies on the location of the microlens array related to the main lens. In Plenoptic 1.0 the microlens array is on the image plane of the main lens ($1/a + 1/b = 1/f$ where f is the focal length of the microlenses) (Georgiev and Lumsdaine, 2009).

4.3 The Camera

In a standard 2D camera, the image is formed by the main lens which projects the image of the scene onto the sensor. In a light field camera, there is an additional array of microlenses. The captures are lens-grid-based representations of the light field. As of today, there are two types of cameras: The Plenoptic 1.0 cameras and the Plenoptic 2.0 cameras (Georgiev and Lumsdaine, 2009) (Figure 4). In the Plenoptic 1.0 camera, the main lens projects the image into the array of microlenses, which then form a set of micro-images on the sensor. The 1.0 approach is characterized by a very simple relationship between the coordinate on the sensor and the light field (x, y, u, v) coordinates, and the resulting reconstructed image has a number of pixels equal to the number of microlenses in the microlens array. An example of a commercial Plenoptic 1.0 camera is the Lytro Camera. In a Plenoptic 2.0 camera, the image formed by the main lens is either in front or behind the microlens array. This approach allows for better resolution, but the price to pay is a complex relationship between the light field (x, y, u, v) and the sensor coordinates. The Raytrix R5 camera[8] is a Plenoptic 2.0 camera (Perwass and Wietzke, 2012) composed of an array of around 7900 microlenses and the image formed by the main lens (called here the *virtual image*) falls behind the microlens array.

It has an additional *extended depth-of-field* property by incorporating three types of microlens with three different focal lengths (Perwass and Wietzke, 2012). The microlenses lie on a hexagonal grid that optimizes the sensor coverage (Figure 5). A raw light field image is composed of a bubble-like pattern; each bubble is the projection of the virtual image by a single microlens. In the rest of the article, a bubble region is called a micro-image, referring to a microlens.

[8]http://www.raytrix.de/tl_files/downloads/R5.pdf

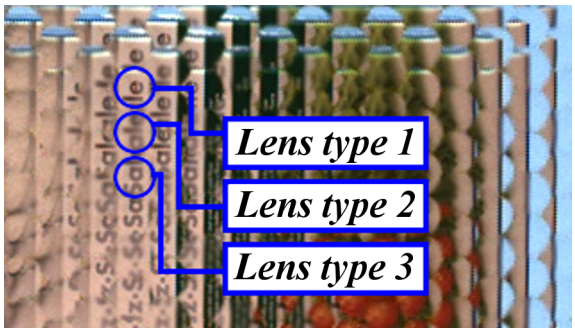

Figure 5: Raytrix technology : Three types of microlenses on a hexagonal grid.

4.4 Possible Usage of The Dataset

The dataset can be used for object classification: it has several classes of objects (5 instances per class) and 6048 captures per object so as to train a detector, which should be able to achieve excellent performance. The 3D pose and the camera-to-object distance are varied and enable testing for scale and pose invariance. The various backgrounds make the dataset interesting for image matting and background subtraction as in (Cho et al., 2014). The range of angles makes it attractive for 3D reconstruction, similar to light field captures done with a camera travelling on a circular rail. Similarly to (Zobel et al., 2002), the dataset can be employed in the domain of object tracking and object pose estimation and prediction. The accuracy of the (relative) poses given by stage precision can be used to assess the accuracy of pose estimators, as well as the accuracy and consistency of corner detectors based on light field data. Eventually the variety of distances can be used in domains requiring scale-invariance.

5 OUR SCALE-INVARIANT RECOGNITION SYSTEM APPROACH

Today's light field object classification approaches require pre-processing like the epipolar-plane images (EPI) (Ghasemi et al., 2014; Ghasemi and Vetterli, 2014). As shown in Figure 6, the image recorded by the plenoptic camera is a group of micro-images lying on a hexagonal grid, each micro-image being the subset of the virtual image of the scene formed by the main lens. The presence of redundancy allows depth estimation (Perwass and Wietzke, 2012).

When we compare two images of the same object with known dimensions taken at two different dis-

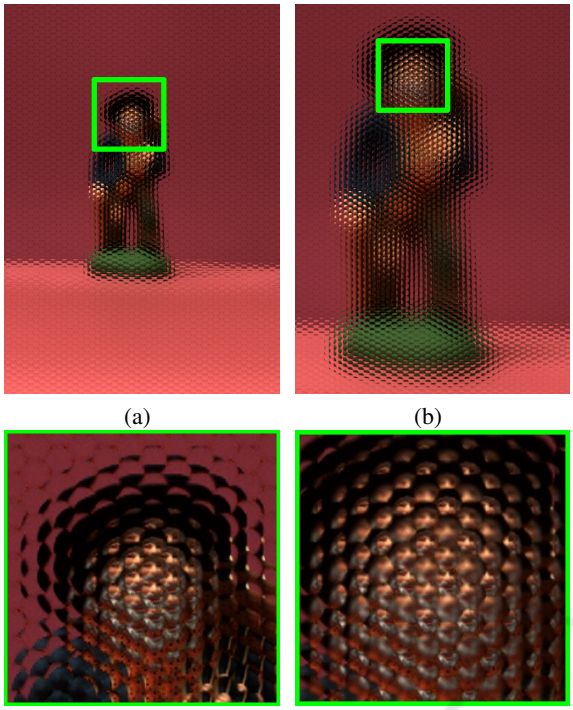

Figure 6: Captures of a figurine at two distances. On a unified background, at the furthest distance from the camera (a) with the corresponding zoomed-in part (b); respectively at the closest distance (c) and (d).

tances, the pattern redundancy is more important on the closest capture than on furthest away one (Figure 6). The existence of three types of microlenses, i.e. three different focal lengths, allows the scene to always be in focus behind at least one type of microlens. We also notice that the scale change within a micro-image is very low. In our approach, we therefore aim at taking advantage of these pattern repetitions and the low scale variation within the micro-images in order to develop a recognition system that is invariant to the scale induced by the distance. Works based on bag-of-visual-words showing interesting performance on object recognition, the underlying intuition of our approach is the counting of patches belonging to a defined dictionary (built at training). The collection of counters results in a unique signature of the object. We expect that the shape of this signature remains the same at any distance, only its amplitude varying with scale (the closer the object, the higher the amplitude).

We propose a baseline method of object recognition from light field imaging to which future works can be compared. It is based on the bag-of-visual-word strategy: (i) a codebook is built from an unsupervised clustering method and (ii) allows building a histogram of each image, the histogram of the test images being compared to the histogram of the trained image that defines the object. The advantage of such a strategy is that histogram computation and comparison are fast to process. Our approach is thus appropriate for real-time implementations.

5.1 Codebook Learning

The codebook is a set of whitened pixel patches learnt from small patches extracted within each micro-image of a training-image set. From a training set made of the raw light field captures of all the objects at the closest distance from the plenoptic camera (one capture per object), we extract n-pixel patches within each micro-image that contains a large part of the object with a defined scanning stride on both axes. For each training raw image, the set of micro-images is extracted from the mask defining the presence of a part of the object. This allows to reduce the noise that patches belonging to the background can introduce in the training process.

The resulting set of patches $x^{(i)} \in \Re^n, i = 1, ..., m$, forms the training vectors. These patches are preprocessed by removing the mean before a PCA-whitening. The final codebook is then learnt by alternating K-means clustering and a cluster-merging step:

Algorithm 1: K-means clustering algorithm. K is the number of clusters, c_i the center of the ith cluster, $i = 1, ..., K$ and N the number of iterations.

1: Initialize the centers c_i of the K clusters from the data
2: Attribute the closest center to each data sample
3: Update each c_i with the mean of all its belonging data sample
4: Repeat N times

5.2 Histogram Extraction and Classification

The codebook is learnt from a set of objects captured at the closest distance and tested on captures at farther distances. Once the codebook learnt, we obtained K centers $c_i \in \Re^n, i = 1, ..., K$. We then extract a normalized histogram of each object containing as many bins as centers.

The test phase is applied within a fixed test region of interest and without any image re-scaling. As a result, the test ROI is composed of micro-images of background and part of the object to recognise. Each n-pixel patch is extracted within each micro-image of the fixed ROI with strides greater than at training phase (Figure 7). The patches can then either belong

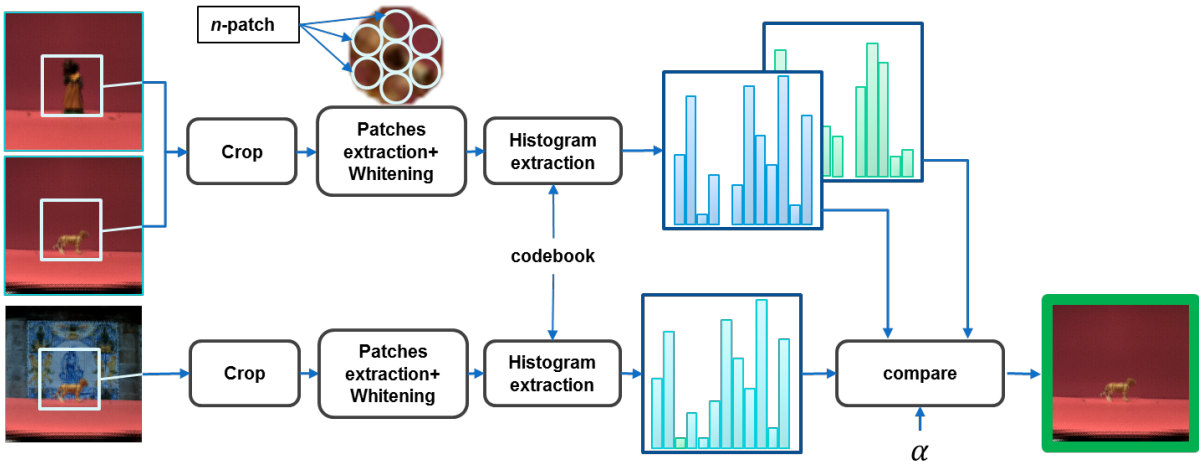

Figure 7: Block diagram of the test scheme. Within a region of interest, each micro-image is used to extract non-overlapping n-patches. The n-patches are PCA-projected according to the training and attributed the closest K-means cluster for building the histogram.

to the object or to the background. After being PCA-whitened, they are attributed a cluster. Each bin of the histogram represents the number of occurrence of the corresponding code in the region of interest. For far object captures, the number of visual words belonging to the object is smaller than for close objects. We expect the histogram of a test image to have its amplitude lower than the one of the training image. We thus scale up the test histogram and compare it with the histogram of the training image of the same object captured at the closest distance. The histograms are compared by minimizing the following thresholded L_1 distance,

$$\arg\min_{\alpha} \sum_{i=1}^{K} |H_T(i) - \alpha * H_t(i)| \qquad (2)$$

where K is the number of histogram bins, α a scaling factor, H_T and H_t the histogram of respectively the training image and the test image. The construction of the histograms and their comparison make our approach compatible with a real-time implementation, the execution time being related to the number of K-mean centers.

6 EXPERIMENTAL RESULTS

We evaluate our scale-invariant object recognition approach on five objects, one instance from each class of our light field dataset, CSEM-25, and with a unique angle. The codebook is trained on the objects located at 28 cm from the camera with the implementation of K-means from (Arthur and Vassilvitskii, 2007). The training set is made of as many images as the number of objects to recognise i.e. five in our evaluation. These images are the ones where the background is uniform and we test on the four subsets, the two firsts with a uniform background and the two seconds with random landscape backgrounds. The training sub-patches are extracted from objects segmented according to the mask provided by the dataset. While the histogram of trained samples is built from the segmented region, the test is applied by extracting the histogram from a fixed window for all the 21 distances (from 28 to 50 cm).

For the experiments, we set a number of clusters to 80 and from 100 to 900, with a step of 200. At training phase, a sub-patch is discarded from the training samples when 25 pixels or more belong to the background. The sub-patch extraction within a micro-image is done with a stride of 1 pixel over the x-axis and 2 pixels over the y-axis. The three RGB color channels are extracted to build the training vectors. At test phase, the detection window is kept constant despite the scale change induced by the camera-to-object distance. We expect the background to introduce noise in the histograms making the recognition challenging for large distances.

The experimental results are presented in table 1. The values represent the average on the detection done with the four backgrounds. At the closest distance, we obtain a recognition rate of 100 %, the background not having a large impact. Using a fixed-sized detection window, the farther the objects, the lower the recognition rate is, due to the noise introduced by the background that fills an increasing proportion of the micro-images. With a few number of bins, we exceed 90 % of correct recognition for each tested distance, the recognition rate expectedly decreasing with the distance.

Table 1: Object recognition on light-field raw images. Recognition rate for 21 tested distances and various number of visual words.

Distance	Number of visual words (K)					
	80	100	200	300	400	500
28	1	1	1	1	1	1
28.6	1	1	1	1	1	1
29.2	1	1	1	1	1	1
29.8	0.95	1	1	1	1	1
30.4	0.95	1	1	1	1	0.95
31.1	0.95	0.95	0.95	0.95	1	0.95
31.9	1	1	1	1	1	0.95
32.7	0.95	1	1	1	0.95	0.95
33.5	0.95	1	1	1	1	0.95
34.4	1	1	1	0.95	0.9	0.9
35.3	1	1	1	0.95	0.9	0.95
36.3	0.95	1	1	1	0.95	0.9
37.4	0.9	1	1	0.95	0.95	0.95
38.6	0.95	1	1	0.9	0.95	0.95
39.8	0.9	0.95	1	0.9	0.9	0.9
41.2	0.95	1	0.75	0.8	0.85	0.85
42.7	0.9	1	0.95	0.8	0.8	0.8
44.3	0.85	0.9	0.95	0.75	0.75	0.7
46	1	1	0.95	0.85	0.8	0.75
47.9	0.95	0.9	0.75	0.65	0.6	0.6
50	1	1	0.8	0.6	0.5	0.45

The results also show an overfitting effect when building a codebook with too many visual words. While the recognition rate remains high for distances close the one of the training set, it drops drastically for farther distances, the cluster means being dependent on the scale of training set of patches.

For a comparative evaluation of our scale-invariant object recognition approach, we tested the corresponding all-in-focus images with a bag-of-feature-based object classifier. As the invariance in scale is our main concern, the SIFT features were employed. After the detection of SIFT keypoints, the SIFT descriptors belonging to the objects were extracted from the all-in-focus training images and clustered to get a k-mean codebook that allows the building of histograms to describe an image. Each all-in-focus test image histogram was compared to the all-in-focus training image histograms with the L2-norm. Figure 8 shows the evaluation results. The experiments were carried out on four and five objects, the fifth object being the fruit. Indeed, we noticed that very few keypoints were extracted from the last object that is not textured enough. The performance of the BOF-based classifier over the five objects are therefore penalized, the fruit not being predictable at test

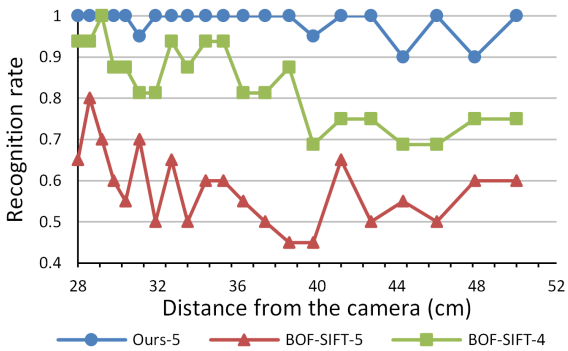

Figure 8: Comparative results showing the performance of our method on light-field images (best results with $K = 100$) and of a bag-of-feature-based approach on standard 2D images (best results with $K = 60$ for 5 objects, BOF-SIFT-5, and $K = 40$ for 4 objects, BOF-SIFT-4).

phase due to insufficient number of SIFT features.

Our scale-invariant object classifier outperforms a 2D BOF-based classifier on three situations: (i) when the distance is known and fixed, (ii) when the scale varies and (iii) on images where scale-space-based features can hardly be extracted.

7 CONCLUSION AND FUTURE WORK

We presented a new approach of scale-invariant object recognition from light field images of our new light field dataset. From our industrial plenoptic camera that has the properties of an extended depth-of-field and micro-images redundancy with low scale variance, we built a real-time recognition system that is robust to large scale variation of almost twice the size of the object when furthest from the camera. With a codebook of a few words (100 visual words) built with a few number of images (one per object to recognise), we reach a recognition rate greater than 90 % despite the scale variation, outperforming a bag-of-feature classifier on standard 2D images. As next steps, we aim at scaling up the system to recognise more objects and also to classify object by category. The dataset is available for download at http://www.csem.ch/csem-25-db.

ACKNOWLEDGEMENTS

The authors thank Pierre-Alain Beuchat, Silvia Marcu and Elio Abi Karam for their assistance with the acquisition set-up automatization, Dr. Amina Chebira for the constructive discussions on the algorithms and

Dr. Guido Bologna for his helpful guidance in the project. This work is supported by the Swiss Hasler Foundation SmartWorld Program, grant Nr. 11083.

REFERENCES

Adelson, E. H. and Bergen, J. R. (1991). The plenoptic function and the elements of early vision. *Computational models of visual processing*, 1(2):2–20.

Arthur, D. and Vassilvitskii, S. (2007). k-means++: the advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, 2007*, pages 1027–1035.

Bolles, R. C., Baker, H. H., and Marimont, D. H. (1987). Epipolar-plane image analysis: An approach to determining structure from motion. *International Journal of Computer Vision*, 1(1):7–55.

Cho, D., Kim, S., and Tai, Y.-W. (2014). Consistent matting for light field images. In *Computer Vision–ECCV 2014*, pages 90–104. Springer.

Cloix, S., Weiss, V., Bologna, G., Pun, T., and Hasler, D. (2014). Obstacle and planar object detection using sparse 3d information for a smart walker. In *VISAPP (2)'14*, pages 292–298.

Coates, A., Karpathy, A., and Ng, A. Y. (2012). Emergence of object-selective features in unsupervised feature learning. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 25*, pages 2681–2689. Curran Associates, Inc.

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22.

Everingham, M., Eslami, S., Van Gool, L., Williams, C., Winn, J., and Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136.

Gavrila, D. M. and Munder, S. (2006). Multi-cue Pedestrian Detection and Tracking from a Moving Vehicle. *International Journal of Computer Vision*, 73(1):41–59.

Georgiev, T. G. and Lumsdaine, A. (2009). Resolution in plenoptic cameras. In *Computational Optical Sensing and Imaging*, page CTuB3. Optical Society of America.

Ghasemi, A., Afonso, N. J., and Vetterli, M. (2014). LCAV-31: a dataset for light field object recognition. In *Proceedings of the SPIE*, volume 9020, San Francisco, California, USA. International Society for Optics and Photonics.

Ghasemi, A. and Vetterli, M. (2014). Scale-invariant representation of light field images for object recognition and tracking. In *Proceedings of the SPIE*, volume 9020 of *Proceedings of SPIE*, San Francisco, California, USA. International Society for Optics and Photonics.

Helmer, S. and Lowe, D. (2010). Using stereo for object recognition. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 3121–3127.

Joshi, N., Matusik, W., and Avidan, S. (2006). Natural video matting using camera arrays. *ACM Trans. Graph.*, 25(3):779–786.

Kim, C., Zimmer, H., Pritch, Y., Sorkine-Hornung, A., and Gross, M. H. (2013). Scene reconstruction from high spatio-angular resolution light fields. *ACM Trans. Graph.*, 32(4):73.

Levoy, M. (2011). The (old) stanford light fields archive. http://graphics.stanford.edu/software/lightpack/lifs.html. [Online, accessed 30-March-2015].

Ng, R. (2005). Fourier slice photography. In *ACM Transactions on Graphics (TOG)*, volume 24, pages 735–744. ACM.

Perwass, C. and Wietzke, L. (2012). Single lens 3d-camera with extended depth-of-field. In *IS&T/SPIE Electronic Imaging*, pages 829108–829108. International Society for Optics and Photonics.

Wanner, S., Meister, S., and Goldluecke, B. (2013). Datasets and benchmarks for densely sampled 4d light fields. In *Vision, Modelling and Visualization (VMV)*.

Wetzstein, G. Synthetic light field archive. http://web.media.mit.edu/gordonw/SyntheticLightFields/index.php. [Online, accessed 30-March-2015].

Wilburn, B., Joshi, N., Vaish, V., Talvala, E.-V., Antunez, E., Barth, A., Adams, A., Horowitz, M., and Levoy, M. (2005). High performance imaging using large camera arrays. *ACM Trans. Graph.*, 24(3):765–776.

Zobel, M., Fritz, M., and Scholz, I. (2002). Object tracking and pose estimation using light-field object models. In *Proceedings of the Vision, Modeling, and Visualization Conference 2002 (VMV 2002), Erlangen, Germany, November 20-22, 2002*, pages 371–378.