Scale-Invariant Representation of Light Field Images for Object Recognition and Tracking

Alireza Ghasemi and Martin Vetterli

AudioVisual Communications Laboratory
École Polytechnique Fédérale de Lausanne

ABSTRACT

We propose a scale-invariant feature descriptor for representation of light-field images. The proposed descriptor can significantly improve tasks such as object recognition and tracking on images taken with recently popularized light field cameras.

We test our proposed representation using various light field images of different types, both synthetic and real. Our experiments show very promising results in terms of retaining invariance under various scaling transformations.

Keywords: Feature Extraction, Transform, Scale Invariance, Plenoptic Function, Light Field Imaging

1. INTRODUCTION

A crucially important part of many computer vision tasks is robust extraction and description of a set of features. These tasks include image matching, stitching, object recognition and face detection among others.[5]

A desirable feature descriptor should be invariant to certain transformations such as scaling and rotation.[1] Several feature detection and description methods have been proposed so far which include Harris corner detector,[2] BRIEF[3] and SIFT.[1]

However, achieving true invariance requires certain information of the scene which are mostly lost during the image formation process in traditional image acquisition devices. An image is usually constructed by projecting a three-dimensional real-world scene to a two dimensional plane.[4] This process is obviously irreversible and therefore information such as depth and occlusion are lost during the image formation. Without having such informations, current algorithms rely mostly on color and texture information for description which leads to a some false matches.

Light field cameras[6] have received wide attention in recent years due to the introduction of end-user products such as the Lytro.[7] Light-field analysis has been applied to various problems in computer vision from face recognition[8] to depth estimation.[9] Moreover, successful efforts have been made to bring the light-field technology to the mobile world.[10]

To achieve the goal of perfect scale-invariant light field feature description, we first study the effect of scaling on the internal structure of light fields. We consider horizontal and vertical slices of the light field signal called epipolar planes. We show that each scene point corresponds to a line in an epipolar plane. Further we show that gradients of the epipolar lines are proportional to the depths of their corresponding scene points. Therefore, scaling the scene can be interpreted as shearing the epipolar plane. We exploit these properties in extracting a scale-invariant feature descriptor from a light-field image.

2. THE PROPOSED APPROACH

2.1 The Plenoptic Function and Formation of Light-Fields

The usual notion of an image as one may expect is a two-parameter function of spatial dimensions x and y. Extending this notion, consider a video sequence which a sequence of 2-D images ordered by time. The whole signal is a 3-D shape in the $x-y-t$ space. This is a generalization of the traditional color or grayscale image.

Going even further, we also relax the assumption of having a single-sensor camera and add another parameter for wavelength of the incident light at the position (x_0, y_0) and time (t_0). This leads to having a four dimensional function that is concatenation of video sequences taken at different wavelengths. Each $x-y$ slice of this volume (assuming

Computational Imaging XII, edited by Charles A. Bouman, Ken D. Sauer, Proc. of SPIE-IS&T Electronic Imaging,
SPIE Vol. 9020, 902015 · © 2014 SPIE-IS&T · CCC code: 0277-786X/14/$18 · doi: 10.1117/12.2041105

other parameters are kept constant) refers to a 2D image captured at a specific time under a specific wavelength. Similar argument applies to $x-y-t$ (video signals) and other slices.

A Plenoptic function is a generalization of a two-dimensional image, in which we have as well as spatial coordinates of the image plane, five more dimensions for the time(t), wavelength (λ) and the position of the camera ($\langle V_x, V_y, V_z \rangle$).[11]

The most general form of the Plenoptic function contains seven degrees of freedom (parameters). This is formed by adding three degrees of freedom for the spatial camera position, one for time and one more for the wavelength, to the tradition (x, y) arguments of the image plane. Therefore we have:

$$P = P_7(x, y, V_x, V_y, V_z, t, \lambda). \tag{1}$$

The plenoptic function is highly structured. It contains a significant amount of information about the captured scene, camera setup, and other acquisition parameters.

However, a seven-dimensional signal is very difficult to acquire and process. In practice, for a specific application a subset of dimensions which contain the most relevant information are captured and further processed. Therefore, we usually reduce the number of parameters by introducing constraints on the capture process. For example, we can consider only static scenes, thereby omitting the time index. Moreover we can omit the wavelength by considering single-sensor lenses. We may also enforce restrictions on the camera movements to reduce the number of parameters even more.

A commonly known 3D restriction of the plenoptic function is the $x-y-V_x$ slice. This is known as the Epipolar Plane Image[12] or the light-field.[13] Light-fields have attracted a lot of attention in recent years.[14–17]

We can capture EPIs using either a linear camera array or equivalently (and more practically) by assuming that the scene is static and linearly moving a single camera. This latter setup is more practical since a plenoptic function (more precisely a slice of it) can be captured easily and efficiently using a handheld device's internal camera.

2.2 Regularity of the Plenoptic Function

Some conditions are explicitly and implicitly assumed when studying plenoptic function and its properties. Firstly, we assume the pinhole camera model for each individual camera and Cartesian coordinates in the plenoptic function. Secondly, we neglect the wavelength parameter λ by considering grayscale images or different color channels. Finally, surfaces in the scene are assumed to be opaque Lambertian surfaces to avoid reflection issues.

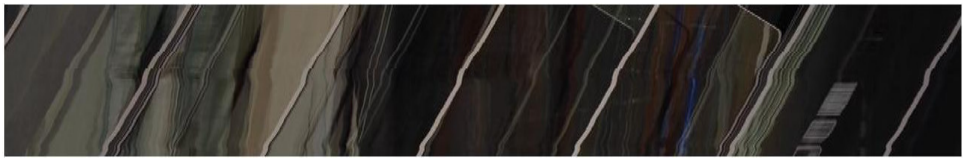
The simplest setup for plenoptic function is what we already know as a still (two-dimensional) image. It can be captured using a single camera and is therefore a sample of the plenoptic function where all parameters except the image coordinates (x and y) are kept constant.

Consider taking a single picture using the pinhole camera model. Assuming that the focal length of the camera is unity and the optical center is located at the origin, the projection of a scene point (X, Y, Z) to the image plane is calculated as:

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \mapsto \begin{pmatrix} \frac{X}{Z} \\ \frac{Y}{Z} \end{pmatrix} \tag{2}$$

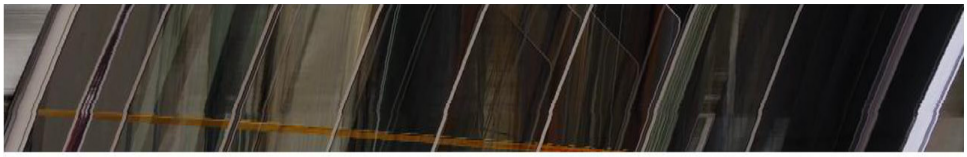
Now assuming the pinhole camera model and Lambertian surfaces,[18] consider an image sequence taken by moving the camera V_x units along the horizontal axis for each image (i.e the light-field or EPI case). Adding a third dimension for the camera location (set to V_x), the mapping from a scene point $P = (X, Y, Z)^T$ to its projection into each image in the sequence can be described as:

(a) Plenoptic Image



(b) Epipolar Plane

(c) Scaled Plenoptic Image



(d) Epipolar Plane in the Scaled Image

Figure 1: Effect of Scaling on the Epipolar Planes

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \mapsto \begin{pmatrix} f\frac{X}{Z} - f\frac{V_x}{Z} \\ f\frac{Y}{Z} \\ V_x \end{pmatrix}. \tag{3}$$

This is how a light-field is formed. We infer from (3) two important facts:

1. Each scene point corresponds to a line (a single-parameter curve) in one of the $x - V_x$ slices of the epipolar volume (the slice corresponding to $y = \frac{Y}{Z}$).
2. The gradient (slope) of this line is proportional to the depth (Z value) of the scene point.

Please note that the above facts are valid only when all our initial assumptions have been met. For example, if the motion is non-linear or speed varies, then the feature paths are no longer lines. An $x - V_x$ slice of an epipolar volume is called an "epipolar plane".

The two fact described above are the starting points for our analysis of the epipolar volume and its properties and how we use it to extract information about the scene.

We will use these properties in deriving our scale-invariant representation for light-field images.

2.3 Scale-Invariance in the Light-Field Domain

The effect of scaling on the plenoptic function can be easily analyzed if we consider the epipolar lines and their properties. We know that the slopes of the epipolar lines are proportional to the depths of their corresponding scene points. Moreover, scaling can be interpreted as adding a constant value to the depth of all points. This is very similar to (but not exactly) rotating the epipolar plane. Figure 1 shows the effect of scaling on the epipolar planes.

Since lines are dominant objects in the epipolar planes, the first step would be to detect them or represent the EPI planes in a new space in which detection of lines is easier. The Radon transform is a well-known approach to achieve this goal.[19] Radon transform R of a function $f(x, y)$ is defined as:

$$R(\rho, \theta) = \int\int f(x, y)\sigma(xCos\theta + ySin\theta - \rho)dxdy \tag{4}$$

In Radon transform, the original image is transformed to a transform plane in which each point (θ_0, ρ_0) corresponds to a line whose equation is $x\,Cos\theta_0 + y\,Sin\theta_0 = \rho_0$ in the original plane. The amount of lightness of the point then determines the length of its corresponding line.

However, implementing the complete Radon Transform is time consuming and also not necessary for our task since the space of possible lines in the EPI plane is limited. The Hough Transform[20] is another line detection approach which is very similar in essence to the Radon's. The main difference lies in the discrete nature of Hough Transform which makes it ideal for our task. There are efficient ways to implement the Hough transform. Algorithm 1 shows the Hough Transform .

Input: The two dimension matrix I representing an image
Discretize the range of θ values in the vector Θ.
Discretize the ρ parameter into n_ρ distinct values.
Construct the $Length(\Theta) \times n_\rho$ output matrix H.
Set all elements of H initially to 0.
foreach *feature point* (x, y) *in* I **do**
  foreach $\theta \in \Theta$ **do**
    $\rho = x \cos\theta + y \sin\theta$
    $H(\theta, \rho) = H(\theta, \rho) + 1$
  end
end
return *the output image* H

Algorithm 1: The Hough Transform Algorithm

The reason that the Hough Transform uses i.e $\theta - \rho$ parameter space rather than the well-known slope-intercept $(m - h)$ form is that both slope and intercept are unbounded even for a finite $x - y$ range that is a digital image.

One key property of the Radon (and Hough) Transform is that rotation in the original plane is converted to a simple translation in the $\theta - \rho$ plane.[21] It is especially useful if we want to derive a rotation-invariant representation from the $\theta - \rho$ image.[22] We can easily do this by setting the minimum θ value to 0. This cancels the effect of any uniform change of θ.

However, what happens during scaling in the epipolar planes is a constant change of m, not θ.[4] We have to cancel the effect of adding a constant Δ to all m_i slopes of epipolar lines. This is not easy to do with the Hough's $\theta - \rho$ space since $\theta = -arccot(m)$ and the relation between $arccot(m)$ and $arccot(m + \Delta)$ is not linear or well-posed.

Therefore, we seek a novel parameterization for representation of lines that, as well as having bounded parameter set, gives us the possibility to isolate uniform changes of slope.

The following exponential parametric form which we hereafter call $\lambda - \mu$ parameterization, has all the desired properties we are seeking:

$$\lambda y + \lambda \ln \lambda x = \mu. \tag{5}$$

In (5), λ is the compressed slope and is calculated from the slope as $\lambda = e^{-m}$. Therefore, for all positive slope values, λ is unique and bounded between 0 and 1. Moreover, μ is also bounded for a bounded range of x and y.

The $\lambda - \mu$ parametrization is not as general as the Hough Transform. First, it can transform only lines with positive slope. Moreover, it is unable to detect fully vertical lines. However, these drawbacks are not important since these situations will not happen in an epipolar plane. The slopes of all lines in an epipolar plane are either positive or negative. Moreover, a vertical line corresponds to a point in infinity which is physically impossible. Therefore, the $\lambda - \mu$ parametrization is sufficient to represent all lines present in an epipolar plane.

The key property of the novel $\lambda - \mu$ parametrization that makes it useful and beneficial for our task is that it can easily isolate and cancel the effect of uniform changes of slope. Suppose λ_m is the λ parameter corresponding to a line with slope m. We observe that

Input: The two dimension matrix I representing an image.
Discretize the range of λ values in the vector Λ.
Discretize the μ parameter into n_μ distinct values.
Construct the $Length(\Lambda) \times n_\mu$ output matrix H.
Set all elements of H initially to 0. **foreach** *each feature point* (x, y) *in* I **do**
  foreach *each* $\lambda \in \Lambda$ **do**
    $\mu = \lambda y + \lambda \ln \lambda x$
    $H(\lambda, \mu) = H(\lambda, \mu) + 1$
  end
end
return *the output image* H

Algorithm 2: The Algorithm for Computing $\lambda - \mu$ Plane (The Exponential Radon Transform)

$$\lambda_{m+\Delta} = \lambda_m e^{-\Delta}. \tag{6}$$

Therefore, the effect of adding a constant Δ to all m_i values can be canceled out by dividing all their corresponding λ_is by the largest one.

2.4 Extracting a Scale-Invariant Feature Vector

We showed that the effect of scaling can be modeled by transforming the $\lambda - \mu$ image as $(\lambda, \mu) \rightarrow (\alpha\lambda, \mu)$. One simple and computationally efficient way to achieve a scale-invariant representation is by integrating the $\lambda - \mu$ plane over lines of equal intercept since only slope is changed at the time of scaling.[23] We define g as

$$g(s) = \int E(\lambda, \mu)\delta(\mu - s\lambda) d\lambda d\mu \tag{7}$$

Where E is the proposed exponential transform.

The use of such method can be justified intuitively. Changing the distance of the camera to the scene, only changes the gradients (slopes) of epipolar lines. Therefore, by integrating over lines of equal intercept, we sum over all possible shearings (slope changes).

The integration approach has the significant benefit that relaxes the need to explicitly extract parameters of the lines from the transform plane. This task requires putting thresholds on the intensities of points in the transform image. Tuning these parameters is itself a time-consuming and complicated task.

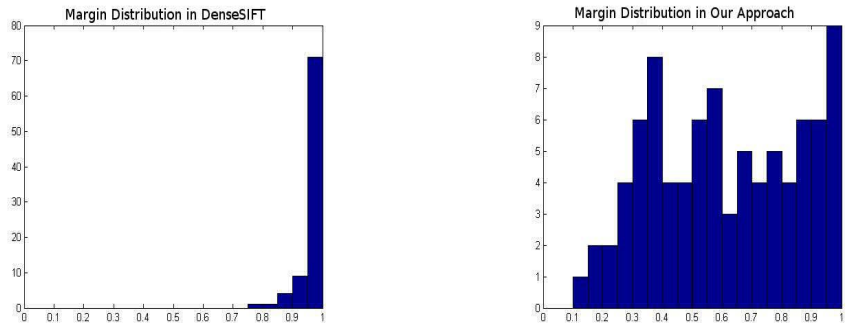
3. EXPERIMENTS AND RESULTS

For our experiments, we used a dataset built by capturing light-field images of multiple buildings in the campus of the École Polytechnique Fédérale de Lausanne (EPFL). The overall dataset is composed of 240 light-field images of 10 buildings in the EPFL campus.

We extracted a single feature vector from each light field by applying the integration (equation 7) to each EPI plane and then applying PCA[24] to extract a subset of the whole concatenated feature vector.

We compared our algorithm with DenseSIFT,[25] which is a recent version of the SIFT descriptor claiming to be faster than it. We also implemented ans tested the Histogram of Oriented Gradients (HOG) approach[26] which has proven successful for outdoor classification tasks.

For the classification task, we used a simple nearest-neighbor approach so that we can see the effects of the extracted features and not the utilized recognition approach. We used a slight modification of the ℓ_1-norm to compute the distance between the feature vectors.



(a) DenseSIFT (b) The Proposed Approach

Figure 2: Distribution of NN/SNN ratios in the DenseSIFT and the Proposed Approach.

Table 1 depicts the results of the proposed approach compared to the HOG and DenseSIFT. We can observe the significant improvement in the proposed approach which is mostly because of the complex geometry inherent in the buildings.

	Our Approach	*DenseSIFT*	*HOG*
Feature Length	**288**	12800	3100
Accuracy (% of Correct Classifications)	**88**	50	61
Average Query Time Per Frame (s)	0.06	**0.04**	1.8

Table 1: Results of the Proposed Approached Compared to the State-of-the-Art

Table 2 depicts the confusion matrix for the proposed approach. The classes in which the recognition accuracy has been mixed are mostly those that have very similar or overlapping geometry and color.

A useful feature in determining the reliability of recognitions in matching algorithms is the ratio between the distance to the nearest neighbor and the second nearest neighbor (NN/SNN ratio[1]). This ratio is a positive values less than 1. High values of NN/SNN ratio of a test sample are not preferred since they show that the nearest neighbor is not highly distinct from the second nearest and therefore may have been selected by chance. These usually show samples which do not belong to any of the classes.

Figure 2 depicts the distribution of NN/SNN ratios in the DenseSIFT and the proposed approach. We observe that the ratios in the DenseSIFT approach are very high and therefore not reliable whereas in the proposed approach the ratios are distributed much more evenly.

Detected / Correct Class	**AGEPOLY**	**BC ATRIUM**	**BC**	**BP**	**INM**	**INN**	**PSA**	**SV INSIDE**	**SV**	**PPH**
AGEPOLY	0.75	0	0.08	0	0	0	0	0	0	0.2
BC ATRIUM	0	0.70	0	0	0	0	0	0	0	0
BC	0.08	0	0.92	0	0	0	0	0	0	0
BP	0	0	0	0.92	0	0	0	0	0	0.2
INM	0.08	0.15	0	0.08	1	0	0	0	0	0
INN	0	0	0	0	0	0.75	0	0	0.5	0
PSA	0	0	0	0	0	0	1	0	0	0
SV INSIDE	0	0.15	0	0	0	0	0	1	0	0
SV	0	0	0	0	0	0	0	0	0.5	0
PPH	0.08	0	0	0	0	0.25	0	0	0	0.6

Table 2: Confusion Matrix for the proposed algorithm. Class names are the names of teh buildings in teh EPFL campus.

SUMMARY

Achieving perfect scale-invariance is usually not possible using classical color image features. This is mostly because of the fact that a traditional image is a two-dimensional projection of the real world.

In contrast, light field imaging makes available rays from multiple view points and thus encodes depth and occlusion information which are very crucial for true scale-invariance. By studying and exploiting the information content of the light field signal and its very regular structure we came up with a provably efficient solution for extracting scale-invariance feature vector representation for more efficient light field matching and retrieval among various views. Our approach is based on a novel integral transform which maps the pixel intensities to a new space in which the effect of scaling can be easily canceled out by a simple integration.

The experiments we conducted on various real-world light field images verify that the performance of the proposed approach is promising in terms of both accuracy and time-complexity. As a probable future improvement, incorporating invariance to various other transforms such as rotation and translation will make the algorithm far more applicable.

ACKNOWLEDGMENTS

This work has been co-funded by the Committee for Technological Innovations (CTI) in Switzerland.

References

[1] Lowe, D., "Distinctive image features from scale-invariant keypoints," *International journal of computer vision* **60**(2), 91–110 (2004).

[2] Harris, C. and Stephens, M., "A combined corner and edge detector.," in [*Alvey vision conference*], **15**, 50, Manchester, UK (1988).

[3] Calonder, M., Lepetit, V., Strecha, C., and Fua, P., "Brief: Binary robust independent elementary features," *Computer Vision–ECCV 2010* , 778–792 (2010).

[4] Berent, J. and Dragotti, P., "Plenoptic manifolds," *Signal Processing Magazine, IEEE* **24**(6), 34–44 (2007).

[5] Hartley, R., Zisserman, A., and ebrary, I., [*Multiple view geometry in computer vision*], vol. 2, Cambridge Univ Press (2003).

[6] Bishop, T. E. and Favaro, P., "The light field camera: Extended depth of field, aliasing, and superresolution," *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(5), 972–986 (2012).

[7] Georgiev, T., Yu, Z., Lumsdaine, A., and Goma, S., "Lytro camera technology: theory, algorithms, performance analysis," in [*IS&T/SPIE Electronic Imaging*], 86671J–86671J, International Society for Optics and Photonics (2013).

[8] Raghavendra, R., Raja, K. B., Yang, B., and Busch, C., "A novel image fusion scheme for robust multiple face recognition with light-field camera," in [*FUSION*], 722–729, IEEE (2013).

[9] Wanner, S., Straehle, C. N., and Goldluecke, B., "Globally consistent multi-label assignment on the ray space of 4d light fields," in [*CVPR*], 1011–1018, IEEE (2013).

[10] Venkataraman, K., Lelescu, D., DuparrÃl', J., McMahon, A., Molina, G., Chatterjee, P., Mullis, R., and Nayar, S., "Picam: An ultra-thin high performance monolithic camera array," in [*ACM SIGGRAPH 2013 Asia*], ACM (2013).

[11] Adelson, E. H. and Wang, J. Y. A., "Single lens stereo with a plenoptic camera," *IEEE Trans. Pattern Anal. Mach. Intell.* **14**(2), 99–106 (1992).

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

[13] Kim, C., Zimmer, H., Pritch, Y., Sorkine-Hornung, A., and Gross, M., "Scene reconstruction from high spatio-angular resolution light fields," *To appear ACM Trans. Graph.(Proc. SIGGRAPH)* (2013).

[14] Tosic, I., Shroff, S. A., and Berkner, K., "Dictionary learning for incoherent sampling with application to plenoptic imaging," in [*ICASSP*], 1821–1825, IEEE (2013).

[15] Do, M. N., Marchand-Maillet, D., and Vetterli, M., "On the bandwidth of the plenoptic function," *IEEE Transactions on Image Processing* **21**(2), 708–717 (2012).

[16] Bando, Y., Holtzman, H., and Raskar, R., "Near-invariant blur for depth and 2d motion via time-varying light field analysis," *ACM Trans. Graph.* **32**(2), 13 (2013).

[17] Marwah, K., Wetzstein, G., Bando, Y., and Raskar, R., "Compressive light field photography using overcomplete dictionaries and optimized projections," *ACM Trans. Graph.* **32**(4), 46 (2013).

[18] Berent, J. and Dragotti, P., "Segmentation of epipolar-plane image volumes with occlusion and disocclusion competition," in [*Multimedia Signal Processing, 2006 IEEE 8th Workshop on*], 182–185, IEEE (2006).

[19] Helgason, S., [*The radon transform*], vol. 5, Springer (1999).

[20] Illingworth, J. and Kittler, J., "A survey of the hough transform," *Computer vision, graphics, and image processing* **44**(1), 87–116 (1988).

[21] Tabbone, S., Wendling, L., and Salmon, J.-P., "A new shape descriptor defined on the radon transform," *Computer Vision and Image Understanding* **102**(1), 42–51 (2006).

[22] Arodz, T., "Invariant object recognition using radon-based transform," *Computers and Artificial Intelligence* **24**(2), 183–199 (2005).

[23] Paplinski, A. P., "Rotation-invariant categorization of colour images using the radon transform," in [*IJCNN*], 1–6, IEEE (2012).

[24] Bishop, C. and en ligne), S. S., [*Pattern recognition and machine learning*], vol. 4, springer New York (2006).

[25] Vedaldi, A. and Fulkerson, B., "Vlfeat: An open and portable library of computer vision algorithms," in [*Proceedings of the international conference on Multimedia*], 1469–1472, ACM (2010).

[26] Dalal, N. and Triggs, B., "Histograms of oriented gradients for human detection," in [*Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*], **1**, 886–893, IEEE (2005).