Scale Normalized Radial Fourier Transform as a Robust Image Descriptor

Evanthia Mavridou, Manh-Dung Hoang, James L. Crowley, Augustin Lux

Scale Normalized Radial Fourier Transform as a Robust Image Descriptor

Evanthia Mavridou[1,2] Mạnh-Dũng Hoàng[2,3] James L. Crowley[1,2] Augustin Lux[1,2]
[1]Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
[2]Inria Grenoble Rhône-Alpes Research Centre, LIG, F-38000 Grenoble, France
[3]Amadeus, Sophia Antipolis, France
{evanthia.mavridou,james.crowley,augustin.lux}@inria.fr, mang-dung.hoang@amadeus.com

***Abstract*—We present a new visual descriptor that combines a multi-scale Laplacian Profile with a Radial Discrete Fourier Transform. This descriptor exists at every position and scale in an image and provides a local feature vector that is both discriminant and robust to changes in orientation and scale. It has a variable description length, and thus can be easily adapted for a variety of applications, ranging from simple detection tasks on low power computing platforms to complex tasks requiring highly discriminant detectors.**

To demonstrate the discriminant power of this descriptor we employ it in its most compact form to construct a cascade of linear classifiers for detecting people in images. We compare this detector to cascades classifiers constructed using Haar wavelets, Gaussian derivatives and variable size block HOG descriptors. Our experiments show that a cascade with this descriptor performs well against the other three detectors when tested using a common publicly available data set. We examine the stability of the descriptor to changes in image rotation and scaling for different description lengths.

***Keywords*—robust image description; detection; matching;**

I. Introduction

In this paper we describe a visual descriptor with variable vector length that can be adapted to a variety of detection problems. The approach used in building this descriptor is to project the overall appearance of an image neighborhood at multiple scales onto a feature vector based on mathematical functions that are equivariant to changes in rotation and scale. While equivariance is somewhat degraded by sampling, the resulting descriptor provides detection that is robust to changes in scale and rotation. The resulting description expresses the local image neighborhood in a manner that separates appearance from orientation and scale and thus can be used to detect local orientation and scale or to describe local appearance independently of rotation and scale.

We demonstrate the discriminant power of this descriptor using the problem of detecting people in images. This is a difficult task because of the large variety of appearances that can result from variations in pose, clothing and illumination. We employ the proposed descriptor to construct a cascade of linear classifiers for detecting people. We compare the resulting detector to three cascade classifiers constructed using Haar wavelets, Gaussian derivatives and variable block-size HOG. Our experiments show that a cascade with this descriptor outperforms the competing detectors when applied to the INRIA Person dataset [13].

This image descriptor can be used both for detection and for image matching. We explore robustness of this descriptor for keypoint detection under changes in scale and orientation, and compare the results with the SIFT descriptor. The results show that this descriptor can be used to provide robust detection of keypoints under such transformations, comparable to those provided by SIFT.

Chapter II discusses the problems of detection and keypoint matching. Chapter III describes a new approach for description of local appearance in images using Laplacian Profile and the Radial Discrete Fourier Transform. Details of the performance evaluation are presented in Chapter IV for detection, followed by Chapter V with experiments on robustness. Chapter VI provides a summary of the technique.

II. Image Description for Human Detection and Keypoint Matching

In this section, we will review popular techniques for people detection and keypoint matching that can provide a baseline for comparison with the proposed technique. The problem of detecting people in images was chosen for its very difficult nature. The human body can have a large variety of configurations in orientations and appearance and occur with a large variety of backgrounds. The aim of the paper is not to demonstrate that the proposed descriptor is best for people detection, but to demonstrate that it can work for visually-difficult vision problems. We used a cascade classifier to construct our comparisons, and thus our comparison is limited to descriptors that can be used with such a classifier.

The proposed approach provides a very general descriptor that can be used for a large variety of visual tasks. Therefore, the selection of suitable experiments is a more challenging than the evaluation of most other state of the art descriptors that have been designed for a limited set of tasks.

A. People detection

Haar features [28] are widely used for detecting people, faces and other visual classes. These features resemble differences of boxes and can be easily computed at very low computational cost using integral images [34]. Haar features can be unstable when used to detect forms that are not aligned with the rows and columns of the image. Nonetheless, Haar features are widely used for detecting visual classes in real time applications.

Gaussian derivatives have long been popular because they can provide scale and rotation invariant description [15], [26]. A Gaussian Pyramid [12] provides a fast algorithm for creating Gaussian derivatives at multiple scales. Gaussian derivatives are widely used to detect edges [8] and interest points [22], [24] as well as for multidimensional histograms of appear-

ance [32]. They have been used with Log Polar Histograms for face detection [17]. More recently, Ruiz [31] has shown that Gaussian derivatives can be used with a cascade classifier to provide a robust real time detector for faces in images. Low order Gaussian derivatives capture visual structures such as bars, blobs and corners, while higher order derivatives can be useful for more complicated patterns but tend to be sensitive to image noise.

Local histograms of image derivatives [32], [33] provide an effective image description for indexing and recognizing visual classes. This approach has been made popular by Dalal and Triggs [13] under the name Histogram of Oriented Gradients (HOG). HOG descriptors are computed as histograms of gradient orientations from a grid of cells. The resulting histograms are concatenated to form a feature vector. Several variants of HOG have been developed to address problems with the original formulation, including the polar-HOG [21] and HOG-LBP [35]. radius and angles

An alternative form of HOG is used to construct a cascade classifier [39]. Zhu et al. employ vectors corresponding to smaller blocks of different sizes and let the cascade training procedure select the most significant ones. They show that although this approach is less discriminant than competing methods, it is faster to compute and can detect a larger variety of human forms. Larger blocks capture information about larger portions of the human form while small blocks cover parts such as legs or arms, providing an improved robustness [5], [39]. Their approach exhibits an accuracy that is similar to the original technique by Dalal and Triggs. We include this technique in our comparisons because of its robustness and lower cost compared to other variants of HOG.

B. Keypoint matching

Invariance to image transformations is very important for keypoint matching. The Scale Invariant Feature Transform (SIFT) descriptor [24], [25] uses local histograms of the orientation of image derivatives over a grid of small windows. Mikolajczyk and Schmid [27] provide a comparison of local descriptors, and show that SIFT like descriptors perform the best, for repeatabililty, followed by Shape Context [4]. More recently, other local descriptors have been proposed including SID [19], LBP-HF [1], ORB [30], LIOP [36], BRISK [20], CARD [3], FREAK [2] and BRIEF [7]. Nevertheless, the most standard comparison test in the literature is with SIFT due to its high state of the art performance and availability of source code. For these reasons, we use SIFT as a baseline to compare the proposed method on rotation and scaling robustness.

III. Creating a new robust descriptor

A. Laplacian profile

Gaussian derivatives can be easily computed as weighted differences of adjacent pixels within a half-octave Gaussian pyramid [6], [10], [12]. The Gaussian filter is a sampled form of a normalized Gaussian function:

$$G(x, y, \sigma) = W_N(x, y)\frac{1}{A}e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (1)$$

where x and y correspond to the integer values of the pixel addresses and σ determines the size (or scale) of visual forms that are described at each image of the pyramid. The term $W_N(x, y)$ corresponds to a window of size $N \times N$ that limits the spacial extent of the sampled Gaussian, where N should be greater than or equal to $8\sigma + 1$. The constant A corresponds to the sum of the coefficients, of the Gaussian. This terms normalizes the gain of the filter to 1, and is necessary to assure a scale invariant impulse response.

The half-octave Gaussian pyramid is composed of K resampled images (pyramid levels), $P(i, j, k)$, each of which has been convolved with a Gaussian filter $G(x, y, 2^{k/2})$ and resampled with a sample distance of

$$s_k = 2^{(k-1)/2} \quad (2)$$

In this paper we will use (x,y) to refer to the original image coordinates of pyramid samples, so that P(x,y,k) is the pyramid sample of level k that corresponds to the pixel (x,y) in the original image. The sampling algorithm for pyramid with $\sqrt{2}$ sampling of the even levels, and the formulae for converting (i,j,k) coordinates to (x,y,k) coordinates may be found in [31].

The number of pyramid levels depends on the size of the original image. For an image composed of H rows of W columns, the pyramid is composed of $K = 2 \times Log_2(min(W, H))$ levels, $P(x, y, k)$, for $k = 1$ to K and will have at most $P = 2 \times N$ samples where $N = W \times H$. The actual number of samples will be slightly smaller, because for the top levels of the pyramid, the impulse response is larger than the original image. Image descriptions at these levels are dominated by boundary effects and can be discarded.

The Laplacian of the image $\nabla^2 p(x, y)$ is the sum of the second derivatives in the row and column. When the image derivatives are computed using Gaussian derivatives, this function exists over a range of Sigmas:

$$LP_{xy}(\sigma_k) =< \nabla^2 G(x, y, \sigma_k), P(x, y) > \quad (3)$$

where "$< -, - >$" refers to the inner product operator. We refer to the function $LP_{xy}(\sigma_k)$ as the "Laplacian Profile" (LP), see figures 1 and 2. The LP is invariant to rotation and exists at every pixel in an image. When computed over a logarithmic scale axis, such as $\sigma_k = 2^{k/2}$, the Laplacian profile is equivariant with image scale [23]. Equivariance refers to the fact that a change in scale of a pattern in an image will result in a shift of the LP along the σ_k axis. Thus a sampled LP provides a rotation invariant feature vector that can be used to recognize patterns independent of scale and also used to determine local characteristic scale.

It is well known that a close approximation to the Laplacian can be provided the difference of samples from adjacent pyramid levels in a half-octave Gaussian pyramid. For each pyramid sample at levels $k = 2$ to K, a Laplacian can be computed by subtracting the pyramid sample at the same image position in level $k - 1$.

$$LP(x, y, k) = P(x, y, k) - P(x, y, k - 1) \quad (4)$$

The samples in this vector can be interpolated to provide a continuous LP for each sample if desired [11].

A Gaussian pyramid composed of $P = 2 \times N$ samples will provide N overlapping LP vectors, with lengths ranging from 1 to $K - 1$. The LP is the spine of the proposed descriptor. The length of the LP defines the number of scales from which a descriptor vector will use information. A LP of five elements takes information from five scales on the pyramid. The length of the LP defines the set of possible heights in the pyramid where the LP can be sampled, e.g. LP of length three can be sampled in a five level pyramid at levels two-three-four and at levels three-four-five.

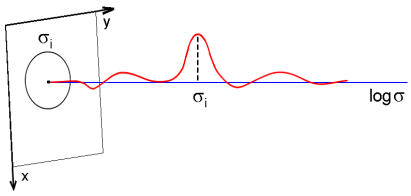

Fig. 1: An example of a LP vector. Laplacian values are collected in a vector at every scale σ.

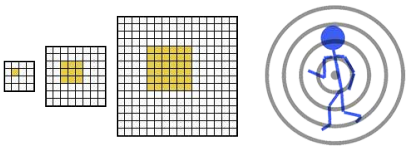

Fig. 2: Left: A sample at a lower resolution corresponds to a larger region in higher resolutions. Its Laplacian inherits this property. Right: A LP vector describes visual appearance simultaneously in several ranges.

These N LP vectors provide a set of variable sized image descriptions. While these descriptors are invariant to rotation and equivariant to scale, they provide only limited discrimination for visual patterns. To make this descriptor more discriminant, we add information from a radial Fourier Transform of a neighborhood around each Laplacian, the size of which can be variable according the requirements of different applications.

B. Radial Discrete Fourier Transform

The most compact version of the proposed descriptor can be acquired by the exploiting the Fourier coefficients of the immediate neighborhoods around a LP. To describe the immediate neighborhood around a Laplacian value, we form a vector from the 4 nearest neighbors from the Laplacian Pyramid for each sample in the LP and perform a Discrete Fourier Transform on these sample. Because these samples are taken from a circle around the LP, this is a Radial Discrete Fourier transform or RDFT. The people detector described below uses RDFT of samples drawn from the Laplacian Pyramid. In the keypoint matching experiments, the RDFT is over samples taken from a Gaussian Pyramid. Our experiments have shown that for a small area, the pixels from Laplacian pyramid are more discriminant.

To provide a description that is equivariant to rotation, we express this RDFT as a magnitude and phase. For the four nearest neighbors Laplacian values, x_0, x_1, x_2 and x_3, around the LP values, we have four coefficients from the 1D DFT: X_0, X_1, X_2 and X_3. The X_0 coefficient corresponds to the average neighborhood value and can be discarded to maintain robustness to illumination intensity. X_1 and X_3 have equal real part and opposite imaginary part, and are thus highly correlated. Thus we use the real and imaginary parts of X_1 to provide a magnitude and phase. The magnitude describes local radial variation for this scale of Laplacian, while the phase gives dominant orientation. So we keep the magnitude that gives a description of appearance that is equivariant to rotation and the phase angle of the X_1 component that tells the dominant orientation of the local neighborhood. We use

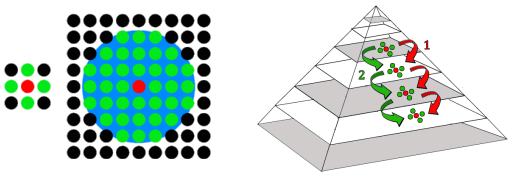

Fig. 3: The central red dots represent the LP. The surrounding green dots represent possible areas for sampling for the RDFT. Left: Four neighbors collected linearly around a LP value for the 1D RDFT (LP-RDFT for cascade) and a disk of samples is taken around a LP value for the 2D RDFT (LP-RDFT for matching). Right: Example of the two steps for extracting a descriptor on an image pyramid. The way is the same for LP-RDFT with either 1D or 2D RDFT.

the sign and absolute value of the X_2 coefficient to provide additional information on appearance. We refer to the final vector as the LP-RDFT.

The LP-RDFT can be extended to cover a larger neighborhood by increasing the radius of the neighbor disk. In this case, the neighborhood disk is mapped onto a 2D grid whose axes are radius and angles. Our experiments show that the resulting descriptor is more discriminant when the samples are drawn from a disc on the original Gaussian pyramid. A 2D DFT is performed over the polar coordinate grid of samples from the Gaussian pyramid. Using polar coordinates is known in image description for achieving rotation invariance [38], [21]. We note that the size of the radii cannot exceed the limits set by the boundaries of the pyramid levels. The resulting 2D RDFT can be expressed using magnitude and phase, where the phase in the rotation direction indicates the dominant direction, and the magnitude signal gives a rotation invariant signature of appearance. Local appearance is best described by lower frequencies while higher frequencies are usually dominated by image noise. Thus we retain only a part of the magnitudes that correspond to lower frequencies. If the size of the 2D RDFT is $M \times N$, then we keep magnitudes from 0 to $M/2$ and 0 to $N/2$. We also normalize the energy (L2-norm) of the resulting vector to obtain robustness to illumination changes.

IV. Experiments on Detection

Two popular approaches for building detectors are Support Vector Machines (SVM) [9] and cascades of linear classifiers [34]. An important disadvantage of SVM is that the trade-off between "false positive" and "false negative" detections is difficult to control. As demonstrated by Viola and Jones [34], Adaboost [16] can be used to construct a cascade of weak linear classifiers for detecting visual patterns that can provide a high true positive detection rate (e.g. 95%) with only a small (e.g. $10^{-7}\%$) false positive rate. In recent years, a number of improvements have been proposed that provide shorter and more effective cascades. For example, the Linear Asymmetric Classifier (LAC) [37], provides a method that respects the asymmetry in the number of positive and negative training images. In our experiments we have used cascade detectors as described in [34] enhanced with LAC to train on the INRIA Person dataset. Our focus has been to compare the effectiveness of our descriptor with Haar wavelets, Gaussian derivative features and variable block-size HOG.

HOG with variable size blocks was chosen because we use a cascade classifier and this descriptor is published as suitable

for a cascade. It is claimed by its authors that this variant of HOG is robust and has less computational cost, though it was still costly for our experiments. We were obliged to use fewer images for this method, considering the fact that the time limit we had to set for training all four cascades is one month and training the cascades with HOG and LP-RDFT occupy almost equally most of this time. In [5] they used fewer images than we did and obtained reasonable results. Though this may sound unfair, the power of this approach is compensated by the qualities of the cascade algorithm which tend to provide a desired false negative and false positive result. In addition, admitting that less images where used for it, actually demonstrates that HOG is a truly strong descriptor that is hard to beat rather that undermining it.

For the LP-RDFT features, we compute a half-octave Gaussian pyramid of the training images and use this to compute a Laplacian pyramid. The training image size 64×128 pixels gives Gaussian pyramids up to seven levels. We discard the highest level because it is dominated by boundary effects and lacks information. We use combinations of the remaining levels to provide LP vectors of different sizes with up to 5 elements. We append neighborhood information at radius 1 pixel for each Laplacian using the magnitude and phase of the X_1 component and the sign and absolute value of X_2 of the RDFT of these neighborhoods to obtain weak classifiers to build our cascade. The final descriptor vectors have up to 25 elements.

The detection cascades for each of the image description methods has a different number of stages. For the Haar cascade, the learning algorithm required 32 stages. For variable size block HOG and the Gaussian derivative cascades, learning resulted in 17 stages. Finally, the LP-RDFT cascade, was constructed using 14 stages. Although the number of levels in each cascade is different, the comparison is fair because we depend on the Receiver Operating Characteristic (ROC) [14] curves to compare the performance of each approach, as seen in figure 4. Testing was performed with the 133 image from the test set that contain only a single human. The ROC curves were made by successively removing stages in the cascade.

We used a 64×128 sliding window with a step of 8 pixels. When a window is identified as positive, it is compared to the groundtruth data and if the overlap at least 60%, it is recorded as a true positive. Otherwise, it is recorded as false positive. We ensure that all windows covering the same person are counted only once. For all except the Gaussian derivatives cascade, we slide the window on the resized image while for the Gaussian derivatives cascade we slide the window on the levels of the Gaussian pyramid of the resized image.

Figure 4 reveals that the Haar feature cascade performs the worst, followed by the Gaussian Derivative cascade. The LP-RDFT cascade and variable size block HOG cascade provide similar results, LP-RDFT providing slightly better performance over most of the range, while variable size block HOG provides better true positive rates at very low false positive rates. The two curves cross on false positive rate 7.873e-05 and detection rate 0.5188. Additionally to the curves, there are two facts that also attest to the superior quality of the proposed descriptor. The one is that the LP-RDFT cascade has up to 14 stages while the variable size block HOG cascade has up to 17 stages. A shorter cascade shows the use of a more powerful description method. The second fact is that the descriptor vectors of the proposed method used in the cascade have variable length up to 25 elements, according to the subset of pyramid levels used, while the variable size block HOG descriptor vectors have always 36 elements [39]. Another important aspect of this figure, is that the ROC of Haar features seems as the less efficient. But taking into account computational time, the Haar features we extract from an image of size 64×128 are computed extremely fast compared to the other three methods. In fact it takes 0.04 sec on a Ubuntu 10.04 LTS computer with 11.8 GiB RAM and Intel Xeon dual core hyperthreading CPU where every processor unit works with 2.33GHz. On the same computer the proposed features for the same image need 0.1 sec to be computed.

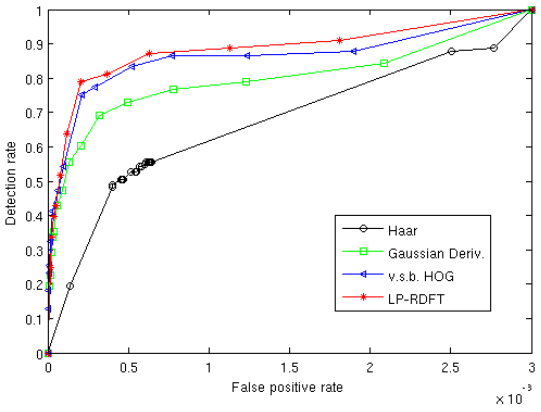

Fig. 4: ROC for the detection task. This figure shows that the LP-RDFT performs better than the other methods. While the improvement over HOG is relatively small, the LP-RDFT offfers an additional advantage in that it relies on a more compact description vector, and results in a detector with fewer stages.

V. Testing robustness on rotation and scaling

A common measure to test robustness for image descriptors is the repeatability of matched keypoints. This measure is defined as the ratio of correctly matched keypoints to the mean of keypoints found in both images [18]. We will use this measure to test different possible versions of our descriptor in order to show how repeatability is affected by the neighborhood size (radius) of the Radial Discrete Fourier Transform. For each of the two image transformations that we test, we take the average results for a set of images.We compare this the repeatability of the SIFT detector for the same task, noting that SIFT is computed using a different form of DoG pyramid. For SIFT we used the usual method for keypoint detection based on maxima in the Difference of Gaussian from three smoothed images with the same sample size. Our descriptor uses a slightly different form of Laplacian pyramid, in which successive layers are resampled by $\sqrt{2}$ by eliminating every second sample [31].

For the proposed descriptor, we test different description lengths ranging from LP vectors without RDFT to descriptors made with larger neighborhoods; we test the descriptor vector as it is used in the cascade in the previous experiments, with the 1D RDFT of the four closest neighbors at radius 1 pixel and descriptor vectors for 2D RDFT at radii of 2, 4, 8 and 10 pixels. For rotation, the descriptor vectors were created by using as many levels as possible from the pyramid with respect to the limits set by the pyramid boundaries and the radii sizes for the RDFT neighborhoods. Therefore, for a particular

radius the descriptor vectors' length is fixed to the maximum possible given the created pyramid, using as much information as possible. For scaling, the vectors were created from a fixed number of five pyramid levels starting on different scales on the pyramid, again with respect to the limits set by the pyramid boundaries and the radii sizes. This way, descriptors sampled on different scales of two pyramids can be matched, making the technique invariant to scale.

For proving the rotation and scale robustness of the proposed method, we use two datasets. For rotation, we use 50 normalized frontal face images of size 128×128 from the FERET face dataset [29]. We selected this dataset because the images are rectangular which helps to determine a circular area of radius 50 pixels within which the matching takes place. This way the keypoints found in the corners of the images which are lost due to rotation are not taken into account. Their small size, 128×128, allows us to test many of them so as to have more accurate results. These images were rotated every 15° around the circle until 360° and with rotation center the center of the image. For scaling, we use another testset with larger images that can be reduced in scale for a larger range of scales. We use four images from the Affine Covariant Features test dataset [27], which are the first image from subsets graf, bikes, ubc and boat. These images were scaled to half their size, four times smaller to their size and eight times smaller to their size. Each time matching was performed between the original and the transformed image and the set of images was always fixed. We use SIFT from the VLFeat open source library and the calculation time for one SIFT feature is less than 1 ms. All test were run on the same computer mentioned at the end of the previous subsection.

Figure 5 show the results for rotation. The averaged curves of the 50 face images for rotation show that the LP alone is relatively robust with rotation. The version of LP-RDFT used in the cascade of the previous chapter, performs poorly under rotation because of the inclusion of the phase (or orientation) information. The most repeatable version of LP-RDFT under rotation is for the 2D RDFT of a neighborhood with radius 2 pixels. This performs the best against all versions of the proposed descriptor. For a single feature of radius 2, the computation time is around 1 ms. Finally, we can see in the figure for rotation that the proposed method performs significantly better than SIFT for small rotations either clockwise or counterclockwise, while SIFT remains repeatable under larger rotations. This difference can be attributed to the way local maxima in the Laplacian are detected in the DoG in the SIFT compared to the way the Laplacian profile is detected in the uniform $\sqrt{2}$ sampled pyramid with our method.

Figure 6 shows the results for scaling. The version of LP-RDFT used in the cascade of the previous chapter does not perform well for scaling. We observe that a larger the radius of the RDFT provides better performance of the LP-RDFT. LP-RDFT with a 2D RDFT of radius 10 performs the best under changes in image scale followed that of radius 8. For a single feature of radius 10, the computation time is around 38 ms and for a feature of radius 8 the time is around 23 ms. Unfortunatelly, these times are too much. The peaks of the plots of LP-RDFT, when the image size correspond to scale factors where the image scale factor coincides with the scale of a level in the image pyramid. We note that the number of detected keypoints decreases with scale, resulting in a decrease

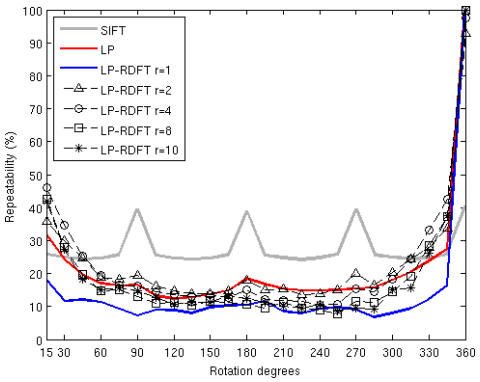

Fig. 5: Repeatability with image rotation. The letter r stands for the neighborhood radius for the RDFT. We note that SIFT provides the most repeatable description, followed by the LP and LP-RDFT with a 2D RDFT of radius 2. The plots are clearer in color.

TABLE I: Computational time for the different versions of the local descriptors. The time is for the computation of one descriptor vector. For the rotation tests, LP-RDFT uses 7 pyramid levels for radius 2 and radius 4, and 5 levels for radius 8 and radius 10. For scaling tests, LP-RDFT uses 5 levels for all radii. Below, the letter r stands for radius used for the RDFT and k for the used pyramid levels (LP length).

SIFT	$<$ 1 ms
LP	$<$ 1 ms
LP-RDFT r=1, k=5	1 ms
LP-RDFT r=1, k=7	1 ms
LP-RDFT r=2, k=5	2 ms
LP-RDFT r=2, k=7	3 ms
LP-RDFT r=4, k=5	9 ms
LP-RDFT r=4, k=7	13 ms
LP-RDFT r=8, k=5	23 ms
LP-RDFT r=10, k=5	38 ms

in repeatability score. This is because there are fewer keypoints to detect in smaller (lower resolution) images.

As it is described in [18], the suitability of a method depends on the requirements of an application. The high repeatability under small changes in rotation and certain scales can be exploited in applications where the parameters of data collection is controlled. Additionally, applications usually need a small number of correspondences to work compared to the number of the correspondences than can be found. LP-RDFT for small radii has a much smaller vector length than SIFT, and is thus much easier to store, and is computed in similar time, providing an efficient alternative to SIFT. On the other hand, if storage and speed are not important, long LP-RDFT vectors can perform provide additional discrimination, and thus provides a trade off of efficiency against computational cost.

With these experiments, we want to demonstrate that though there is space for improvement, robustness to scale and rotation are evident. These shows that the proposed method can compete in detection tasks where dense descriptors are suitable and in the same time be able to be adjusted (without changing its theory) and used as a local descriptor for keypoint matching. Most descriptors in order to be adjusted to very different tasks, usually need extensive alterations to their theoretical basis.

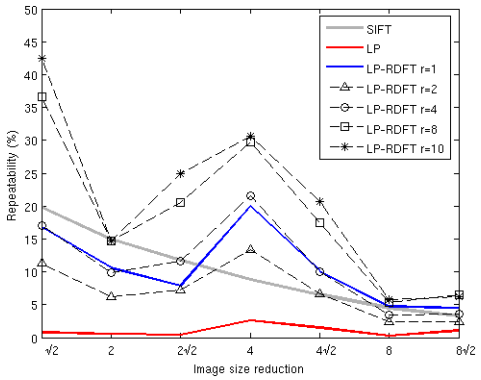

Fig. 6: Repeatability under changes in image scale. The horizontal axis shows the scale factor for smaller scaled images. The letter r stands for the neighborhood radius for the RDFT. LP-RDFT with a 2D RDFT of radius 10 is the most repeatable under changes in image scale followed by LP-RDFT with a 2D RDFT of radius 8. Relative peaks appear when the scale factor of the test data corresponds to the scale of the pyramid levels. The plots are clearer in color.

VI. Conclusions

In this paper we have described a new visual descriptor composed of Laplacian Profiles augmented with a Radial Discrete Fourier Transform. We have evaluated the discriminant power of this descriptor using the problem of people detection in images and explored its robustness on keypoint matching. The experiments showed that this descriptor can provide state of the art performance on detection with significantly smaller description length while providing robustness and equivariance to changes in rotation and scale.

More generally, this descriptor is an example of a new class of image descriptors that combine the scale invariance of a Laplacian Profile with the rotation invariance of a Radial Fourier Transform, and provide a description of local image neighborhoods that can be made robust and adjustable to a variety of applications.

References

[1] T. Ahonen, J. Matas, C. He, and M. Pietikäinen. Rotation Invariant Image Description with Local Binary Pattern Histogram Fourier Features. In *SCIA*, 2009.

[2] A. Alahi, R. Ortiz, and P. Vandergheynst. FREAK: Fast Retina Keypoint. In *CVPR*, New York, 2012. Ieee. CVPR 2012 Open Source Award Winner.

[3] M. Ambai and Y. Yoshida. Card: Compact and real-time descriptors. In *ICCV*, pages 97–104, 2011.

[4] S. Belongie and J. Malik. Matching with shape contexts. In *IEEE Workshop on Content-based Access of Image and Video Libraries*, pages 20–26, 2000.

[5] J. Brookshire. Person Following Using Histograms of Oriented Gradients. *International Journal of Social Robotics*, 2:137–146, 2010.

[6] P. Burt and E. Adelson. The laplacian pyramid as a compact image code. IEEE *Trans. on Pattern Analysis and Machine Intelligence*, 31:532–540, 1983.

[7] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua. Brief: Computing a local binary descriptor very fast. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(7):1281–1298, 2012.

[8] J. Canny. A Computational Approach to Edge Detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679 –698, November 1986.

[9] C. Cortes and V. Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.

[10] J. L. Crowley and A. C. Parker. A Representation for Shape Based on Peaks and Ridges in the Difference of Low-Pass Transform. IEEE *Trans. on Pattern Analysis and Machine Intelligence*, 6(2):156–170, March 1984.

[11] J. L. Crowley, O. Riff, and J. H. Piater. Fast Computation of Characteristic Scale Using a Half-Octave Pyramid. In *Scale Space 03: 4th International Conference on Scale-Space theories in Computer Vision, Isle of Skye*, 2002.

[12] J. L. Crowley and R. M. Stern. Fast Computation of the Difference of Low-Pass Transform. IEEE *Trans. on Pattern Analysis and Machine Intelligence*, 6(2):212–222, March 1984.

[13] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[14] T. Fawcett. ROC Graphs: Notes and Practical Considerations for Researchers. Technical report, HP Laboratories, 2004.

[15] W. Freeman and E. Adelson. The design and use of steerable filters. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(9):891 –906, September 1991.

[16] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *EuroCOLT*, 1995.

[17] D. Hall and J. L. Crowley. Face detection by robust generic features computed from luminance. *RFIA*, 2004.

[18] L. Juan and O. Gwon. A Comparison of SIFT, PCA-SIFT and SURF. *International Journal of Image Processing (IJIP)*, 3(4):143–152, 2009.

[19] I. Kokkinos and A. Yuille. Scale Invariance without Scale Selection. In *CVPR*, 2008.

[20] S. Leutenegger, M. Chli, and R. Siegwart. Brisk: Binary robust invariant scalable keypoints. In *ICCV*, pages 2548–2555, 2011.

[21] W. Li, W. Chengdong, C. Dongyue, and L. Baihua. Rotation-Invariant Human Detection Scheme Based on Polar-HOGs Feature and Double Scales Direction Estimation. In *SOPO*, 2011.

[22] T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, 1994.

[23] T. Lindeberg. On the axiomatic foundations of linear scale-space: Combining semi-group structure with causality vs. scale invariance., 1997.

[24] D. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[25] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vision*, 60(2):91–110, November 2004.

[26] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[27] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[28] C. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *ICCV*, 1998.

[29] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET Evaluation Methodology for Face-Recognition Algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(10):1090–1104, October 2000.

[30] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB: An Efficient Alternative to SIFT or SURF. In *ICCV*, 2011.

[31] J. A. Ruiz-Hernandez, A. Lux, and J. L. Crowley. Face detection by cascade of Gaussian derivates classifiers calculated with a Half-Octave Pyramid. In *FG*, 2008.

[32] B. Schiele and J. L. Crowley. Object recognition using multidimensional receptive field histograms. In B. Buxton and R. Cipolla, editors, *Computer Vision ECCV '96*, volume 1064 of *Lecture Notes in Computer Science*, pages 610–619. Springer Berlin Heidelberg, 1996.

[33] B. Schiele and J. L. Crowley. Recognition without Correspondence using Multidimensional Receptive Field Histograms. *International Journal of Computer Vision*, 36:31–50, 2000.

[34] P. Viola and M. Jones. Rapid Object Detection using a Boosted cascade of Simple Features. In *CVPR*, 2001.

[35] X. Wang, T. Han, and S. Yan. An hog-lbp human detector with partial occlusion handling. In *CVPR*, 2009.

[36] Z. Wang, B. Fan, and F. Wu. Local Intensity Order Pattern for feature description. In *ICCV*, 2011.

[37] J. Wu, S. C. Brubaker, M. D. Mullin, and J. M. Rehg. Fast Asymmetric Learning for cascade Face Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(3):369–382, March 2008.

[38] D. Zhang and G. Lu. Generic fourier descriptor for shape-based image retrieval. In *ICME*, 2002.

[39] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan. Fast Human Detection Using a cascade of Histograms of Oriented Gradients. In *CVPR*, 2006.