Histogram-based Interest Point Detectors

Wei-Ting Lee and Hwann-Tzong Chen
Department of Computer Science, National Tsing Hua University
30013 Hsinchu, Taiwan

Abstract

We present a new method for detecting interest points using histogram information. Unlike existing interest point detectors, which measure pixel-wise differences in image intensity, our detectors incorporate histogram-based representations, and thus can find image regions that present a distinct distribution in the neighborhood. The proposed detectors are able to capture large-scale structures and distinctive textured patterns, and exhibit strong invariance to rotation, illumination variation, and blur. The experimental results show that the proposed histogram-based interest point detectors perform particularly well for the tasks of matching textured scenes under blur and illumination changes, in terms of repeatability and distinctiveness. An extension of our method to space-time interest point detection for action classification is also presented.

1. Introduction

Detecting distinctive invariant low-level features in images is a fundamental aspect of many computer vision tasks. The effectiveness of existing local feature detectors has been well demonstrated through varied vision applications, *e.g.*, [8], [4], [16], [23], [27]. Thorough experimental comparisons and performance evaluations on popular interest-point and local-feature detectors are also available in the literature of computer vision [21], [24]. More recently, Tuytelaars and Mikolajczyk have presented an overview of local invariant feature detectors [28]. These evaluations and literature surveys provide a systematic way to gain insight into the characteristics of widely used feature detection methods. It has been shown that interest point detectors such as the Harris corner detector [11] and the Hessian-based interest point detectors [2], [15] are important building blocks of various local invariant feature detectors. For example, the Harris-Affine and Hessian-Affine detectors [18], [19], [21] are based on affine normalization around Harris and Hessian points; the SURF detector [1] relies on the determinant of the Hessian matrix for selecting the location and the scale; the SIFT detector [17] eliminates unstable edge responses by analyzing the Hessian matrix of the intensity surface.

The Harris corner detector [11] is considered to be one of the most reliable interest point detectors [24]. It is popular owing to its robustness to rotation, illumination variation, and image noise. Briefly, the Harris detector uses the second moment matrix, also called the auto-correlation matrix, to explore the local statistics of image intensity variations, with patches shifted by a small amount in different directions. The locations of interest points can be identified by analyzing the trace and the determinant of the second moment matrix, derived from first-order derivatives of image intensity function. The second moment matrix can be directly extended to RGB color space by combining the three color channels [22]. The underlying idea of the Hessian-based detectors is similar to the Harris corner detector. The Hessian detector explores the second-order Taylor expansion of the (Gaussian convolved) intensity surface, and the resulting Hessian matrix that consists of the second-order derivatives describes the local image structures. Similarly, the trace and the determinant of the Hessian matrix can also be used to decide the interest points. More detailed discussions on interest point detectors and local feature detectors can be found in [21], [24], [28].

This paper presents a new approach to the detection of interest points using histogram information. The proposed detectors are able to identify interest points that exhibit a distinctive distribution of low-level features in a local area. Whereas histogram-based representations have been widely used by the feature descriptors, *e.g.*, HOG [6], SIFT [17], GLOH [20], existing interest point detectors simply use pixel-based (intensity or color) representations to characterize local features. For images consisting of highly textured objects such as brick walls or trees, using pixel-based information to detect interest points may yield too many responses of less stable corners—Shifting a textured patch by a small amount may cause a significant increase in the sum of squared differences, even though the variations in the distributions of texture patterns should be insignificant. Moreover, since popular descriptors for matching or object recognition are built from histograms of low-level features, the feature descriptions extracted from the locations of Har-

ris or Hessian points may not be distinctive enough to distinguish the detected regions having similar distributions of textured patterns.

One of the aim of this paper is to bridge the gap between local feature detectors and descriptors. We incorporate histogram-based representations into the detection process of interest points. The proposed interest point detectors have strong invariance to rotation, illumination variation, and blur. Scale invariance can be achieved by searching for stable points across several possible scales. Kadir *et al*. have presented a salient point detector that also takes account of local intensity histograms [12], [13]. Their algorithm uses the intensity histogram to measure the saliency by local entropy, and seeks to extract local regions that have high complexity. However, local entropy might not be a sufficient criterion to find stable points for matching, as shown through the experiments in [21]. Furthermore, for a textured image that exhibits high entropy values almost everywhere, a region of low entropy in the image should be considered salient as well. Dorkó and Schmid [7] introduce description stability as a criterion for scale selection: They use the common Harris and Laplacian detectors to find interest points, and at each location of detected interest point, the SIFT descriptors under various scales are computed. The scale for which the change of description is minimal will be chosen. Our detectors, on the other hand, are directly built on the histogram-based representations and similarity measures, and thus do not need to compute the SIFT descriptions in advance. The experimental results show that the histogram-based interest point detectors perform well for image matching tasks, especially in matching textured scenes under blur and illumination changes.

2. Detecting Histogram-based Interest Points

For each pixel (x_i, y_i) in a given image patch, we may derive a discrete quantity $b(x_i, y_i)$ from the low-level image features such as color or oriented gradient. Assume that the discrete quantity has L levels (bins) in the interval $[1, L]$. The function $b : \mathbb{R}^2 \rightarrow \{1, \ldots, L\}$ thus associates to the pixel (x_i, y_i) the index $b(x_i, y_i)$ of the histogram bin corresponding to the low-level image feature of that pixel.

At each pixel location (x, y), the value of the kth bin of a weighted histogram $h(x, y) = \{h_k(x, y)\}_{k=1, \ldots, L}$ is given by

$$h_k(x, y) = \frac{1}{Z} \sum_{\substack{(x_i, y_i) \\ \in \Omega(x,y)}} w(x_i - x, y_i - y) \, \mathbf{1}_{\{b(x_i, y_i) = k\}}, \quad (1)$$

where $w(x, y) = e^{-(x^2 + y^2)/2\sigma^2}$ is a Gaussian weighting function, and $\mathbf{1}_{\{\cdot\}}$ is the indicator function. The set $\Omega(x, y)$ defines a neighborhood around (x, y), and Z is a normalization term to ensure $\sum_{k=1}^{L} h_k(x, y) = 1$. Note that the value

of σ in $w(x, y)$ and the size of the neighborhood Ω depend on the chosen scale of interest points to be detected.

Given a shift $(\Delta x, \Delta y)$ and a pixel (x, y), we may use the Bhattacharyya coefficient [3], [5] to measure the similarity between the histogram $h(x, y)$ around (x, y) and the histogram $h(x + \Delta x, y + \Delta y)$ around the shifted pixel location. The Bhattacharyya coefficient between $h(x, y)$ and $h(x + \Delta x, y + \Delta y)$ is defined by

$$\rho = \sum_{k=1}^{L} \sqrt{h_k(x, y) \, h_k(x + \Delta x, y + \Delta y)}, \quad (2)$$

where $\rho \in [0, 1]$, and $\rho = 1$ if the two histograms are identical. The Bhattacharyya coefficient ρ can be approximated by a Taylor series truncated to the second terms, that is,

$$
\begin{aligned}
\rho \;\approx\; & \frac{1}{2} \sum_{k=1}^{L} \sqrt{h_k(x, y) h_k(x, y)} \\
& + \; \frac{1}{2} \sum_{k=1}^{L} h_k(x + \Delta x, y + \Delta y) \sqrt{\frac{h_k(x, y)}{h_k(x, y)}} \\
& + \; \frac{1}{2} \begin{bmatrix} \Delta x & \Delta y \end{bmatrix} H(x, y) \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \\
\;=\; & 1 + \frac{1}{2} \begin{bmatrix} \Delta x & \Delta y \end{bmatrix} H(x, y) \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix}. \quad (3)
\end{aligned}
$$

A more detailed derivation is given in the appendix. The Hessian matrix $H(x, y)$ in (3) is a 2-by-2 symmetric matrix containing the second-order partial derivatives of the Bhattacharyya coefficient ρ. Since the Bhattacharyya coefficient has a local maximum at $(\Delta x, \Delta y) = (0, 0)$, the Hessian matrix of ρ is negative semidefinite. More specifically, $H(x, y)$ is defined by

$$H(x, y) = \sum_{k=1}^{L} \sqrt{h_k(x, y)} \begin{bmatrix} \frac{\partial^2 \sqrt{h_k(x,y)}}{\partial x^2} & \frac{\partial^2 \sqrt{h_k(x,y)}}{\partial x \partial y} \\ \frac{\partial^2 \sqrt{h_k(x,y)}}{\partial x \partial y} & \frac{\partial^2 \sqrt{h_k(x,y)}}{\partial y^2} \end{bmatrix}. \quad (4)$$

By taking into account the Gaussian weighting function in (1), we obtain the elements of $H(x, y)$ as

$$H_{1,1}(x, y) = \left(\frac{x^2}{\sigma^4} - \frac{2}{\sigma^2} \right) - \frac{1}{Z^2 \sigma^4} \sum_{k=1}^{L} \frac{m_X^2}{h_k(x, y)}, \quad (5)$$

$$H_{2,2}(x, y) = \left(\frac{y^2}{\sigma^4} - \frac{2}{\sigma^2} \right) - \frac{1}{Z^2 \sigma^4} \sum_{k=1}^{L} \frac{m_Y^2}{h_k(x, y)}, \quad (6)$$

$$H_{1,2}(x, y) = H_{2,1}(x, y) = \frac{xy}{\sigma^4} - \frac{1}{Z^2 \sigma^4} \sum_{k=1}^{L} \frac{m_X m_Y}{h_k(x, y)}, \quad (7)$$

where

$$m_X = \sum_{\substack{(x_i, y_i) \\ \in \Omega(x,y)}} x_i \, w(x_i - x, y_i - y) \mathbf{1}_{\{b(x_i, y_i) = k\}}, \quad (8)$$

and

$$m_Y = \sum_{\substack{(x_i,y_i) \\ \in \Omega(x,y)}} y_i\, w(x_i - x, y_i - y) \mathbf{1}_{\{b(x_i,y_i)=k\}}\,. \quad (9)$$

Matrix $H(x,y)$ captures the histogram structure of the local neighborhood around pixel (x,y). If the absolute values of both eigenvalues of $H(x,y)$ are large, then a shift $(\Delta x, \Delta y)$ in any direction will result in a significant drop of the Bhattacharyya coefficient, and therefore, the histogram $h(x,y)$ of some low-level image features around (x,y) should be quite dissimilar to the histograms around neighboring pixels $(x + \Delta x, y + \Delta y)$. We consider such a pixel to be an interest point. The problem of identifying interest points can be handled through observing the eigenvalues of the Hessian matrix corresponding to the local Bhattacharyya coefficient. As shown by Harris and Stephens for the Harris corner detector [11], we also do not need to compute the eigenvalues explicitly. The absolute values of the eigenvalues for Hessian matrix $H(x,y)$ can be modeled by a response function R on the determinant and the trace:

$$R(H) = \det(H) - \kappa \operatorname{trace}^2(H)\,, \quad (10)$$

where we use $\kappa = 0.1$ for the experiments presented in this paper. If a Hessian matrix H has a high response, it is more likely that its both eigenvalues are of large absolute values. The response function is used to decide whether a pixel is an interest point. Non-maximum suppression is applied to the responses of all pixels, and local maxima are selected as nominated interest points.

3. Extracting Local Invariant Regions for Matching

We describe in this section how to apply our interest point detector to the matching tasks that rely on the detection of local invariant regions. We present two possible choices of histogram-based representations for our method, and discuss the process of selecting the scale.

3.1. Histogram-based Image Representations

We may represent an image patch by a histogram of low-level image features. Different types of histogram representations can be incorporated into our method to build the histogram-based interest point detectors. Described below are two types of histograms that are tested in our experiments.

Color Histogram. Color histograms are commonly used in object tracking and image retrieval as the image representation. We employ the color representation proposed by Comaniciu *et al.* [5] to describe local image structures.

We quantize each color channel (256 levels assumed) in RGB color space into 8 bins, and obtain a histogram with $8 \times 8 \times 8 = 512$ bins. The quantization function is given by $b(x,y) = \lfloor R_{x,y}/32 \rfloor \times 8^2 + \lfloor G_{x,y}/32 \rfloor \times 8 + \lfloor B_{x,y}/32 \rfloor + 1$, where $R_{x,y}$, $G_{x,y}$, and $B_{x,y}$ are the RGB values of pixel (x,y). By plugging the function b into the weighted histogram in (1), as well as m_X and m_Y in (8) and (9), we obtain the Hessian matrix $H(x,y)$ in (4) for identifying interest points.

Oriented Gradient Histogram. Intensity gradients can also be used as the low-level features for constructing histograms. We quantize the orientation of gradient into 8 bins, each of which covers a 45-degree angle. The magnitude of gradient is also divided into 8 bins, and thus the resulting histogram contains $8 \times 8 = 64$ bins. Because the magnitude of gradient might provide useful information for describing local regions, the formulation of histogram in (1) can be modified to include the magnitude of gradient:

$$h_k(x,y)$$
$$= \frac{1}{Z} \sum_{\substack{(x_i,y_i) \\ \in \Omega(x,y)}} w(x_i - x, y_i - y)\, \|\mathbf{g}(x_i,y_i)\|^\alpha\, \mathbf{1}_{\{b(x_i,y_i)=k\}}\,,$$
$$(11)$$

where $\|\mathbf{g}(x_i,y_i)\|$ is the magnitude of gradient at pixel (x_i,y_i), and α is a scaling parameter. The equations of m_X and m_Y in (8) and (9) also need to be modified correspondingly to get the Hessian matrix.

3.2. Scale Selection for Detecting Local Invariant Features

Scale issue aside, the proposed histogram-based detection algorithm can be used as a stand-alone interest point detector. Nevertheless, for vision applications such as matching or object recognition, it is critical to find local features that are invariant to scale changes. We may explore the scale space by taking account of σ of the weighted histogram in (1). A small variation $\Delta\sigma$ can be added to σ, and we get an equation of the Bhattacharyya coefficient subject to σ and $\sigma + \Delta\sigma$, similar to the formulation in (2). In our implementation we use a more straightforward approach by detecting interest points at each given scale based on the response $R(H)$ in (10). In addition to the scale, the 'shape' of the local region can also be determined by analyzing the Hessian matrix H. The estimation of scale and shape helps to extract the feature descriptions more faithfully for matching. We show in next section that the local invariant regions selected by our method are effective in matching scenes under blur and illumination variations. Fig. 1 illustrates some examples of detected interest points using color histograms and oriented gradient histograms.
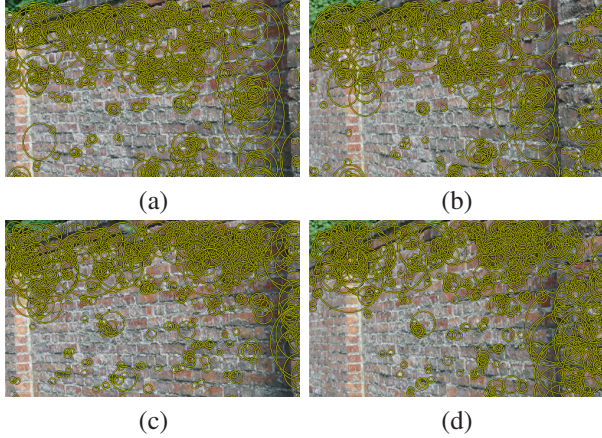
(a)   (b)

(c)   (d)

Figure 1. Detected interest points. (a)&(b) Using color histogram. (c)&(d) Using oriented gradient histogram.

4. Experiments

We apply the proposed histogram-based interest point detectors to the tasks of matching scenes under different imaging conditions. The performance is evaluated by the measures of repeatability and distinctiveness. We also present an extension of our method to the detection of space-time interest points in video sequences, for the application of action classification.

4.1. Image Matching

This experiment is focused on applying our method to the matching problem. We use the test data provided by [21] to evaluate our approach. The test data have eight sets of images, and each set contains six images with five homographies between the first reference image and the other five images. Different variations in imaging conditions for structured and textured scenes are included in the test data: they are viewpoint changes, scale changes, image blur, JPEG compression, and illumination changes.

The two histogram-based detectors presented in Section 3 are performed to find interest points in the test images. For the detector with color histogram, we apply preprocessing to the input images with histogram equalization and Gaussian smoothing for each RGB channel. For the detector based on the oriented gradient histogram, only the intensities of the input images are used. In this experiment, we omit the magnitude term of the oriented gradient histogram in (11) by setting the scaling parameter α to zero. We detect the interest points with eight scales: We have $\{\sigma_d\}_{d=1,\dots,8} = \{2, 2\sqrt{2}, 4, \dots, 16\sqrt{2}\}$ for the weighting function $w(x,y)$ in (1), and the corresponding sizes of the neighborhood are $\{|\Omega_d(x,y)|\}_{d=1,\dots,8} = \{15 \times 15, 21 \times 21, 29 \times 29, \dots, 159 \times 159\}$, where the diameter of neighborhood is computed by $2 \cdot \mathrm{round}(3.5\,\sigma) + 1$. The image

region defined by the neighborhood of an interest point will be used for performance evaluation. In practice, instead of increasing the Gaussian-kernel scale σ and the neighborhood size, we downsample the input images by a step of $\sqrt{2}$ and use a fixed neighborhood size of 15×15 with $\sigma = 2$. This approximation would result in a slight loss of matching accuracy, but could greatly reduce the computational cost.

The evaluation is based on two metrics, the repeatability and the matching score, as described in [21]. To compute the two metrics, we use the Matlab code provided by Mikolajczyk *et al*. [21] (available from http://www.robots.ox.ac.uk/~vgg/research/affine/). A high repeatability score means that the detector can stably find corresponding regions, given by the interest points, in two scenes with some transformation. Apart from the repeatability, for practical applications such as matching or recognition, it would be important to find corresponding regions that produce distinctive feature descriptions. The extracted regions need to be distinguishable from other regions, so that the correct correspondences can be identified through comparing the similarities between their feature descriptions. The distinctiveness of the detected regions is measured by the matching score .

We present the repeatability and the matching scores of our detectors in Fig. 2 and Fig. 3. For comparison, we also include the results of MSER, Harris-Affine, and Hessian-Affine detectors from [21]. Note that, as pointed out in [21], there is no detector which outperforms others in all the experiments. The detector using color histogram generally performs better than the one using oriented gradient histogram. Our results in Fig. 2 are not as good as the state-of-the-art, especially when the scenes present large changes in scale and viewpoint angle. However, our detectors achieve very good performance for the tasks of matching scenes with blur and illumination changes, see Fig. 3. Our detectors perform particularly well for the *Trees* sequence, as shown in Fig. 3(c). The number of correct nearest neighbor matches produced by our detectors usually ranges from 100 to 1000, depending on the image content.

Complexity and Required Computation Time. Assume that the input image contains N pixels, and the histogram has L bins. Let $|\Omega_1|$ denote the size of the smallest neighborhood for computing the histogram with approximation of downsamping. The values of $h_k(x,y)$, m_X, and m_Y are computed by convolution. Thus the complexity is $O(NL \times |\Omega_1|)$ for a single scale. If D different scales are taken into consideration, then we get the final complexity $O(NLD \times |\Omega_1|)$. Given that an input image of size 1000×700 pixels, $L = 512$ bins, $D = 8$ scales, and $|\Omega_1| = 15 \times 15$ pixels, it would take 40 seconds to compute all the required responses of $R(H)$ in Matlab on an Intel Core2 Duo 2.33GHz PC.
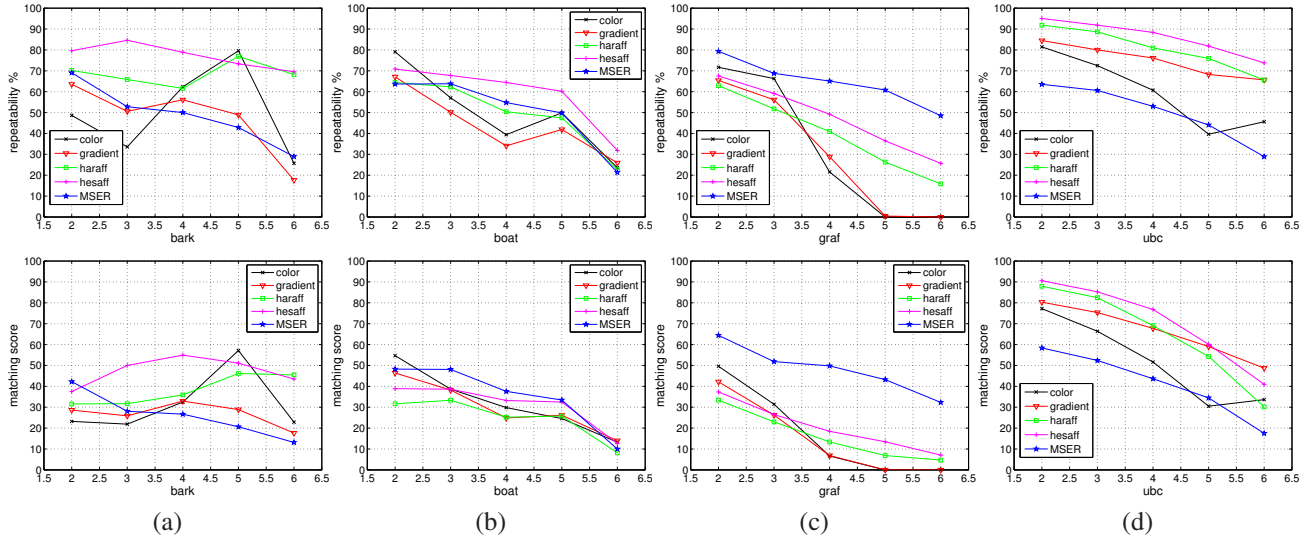
Figure 2. Repeatability and matching score. (a) Scale change for the textured scene. (b) Scale change for the structured scene. (c) Viewpoint change. (d) JPEG compression. Our detectors do not perform very well on these data sets, in comparison with the best results of other popular detectors presented in [21].
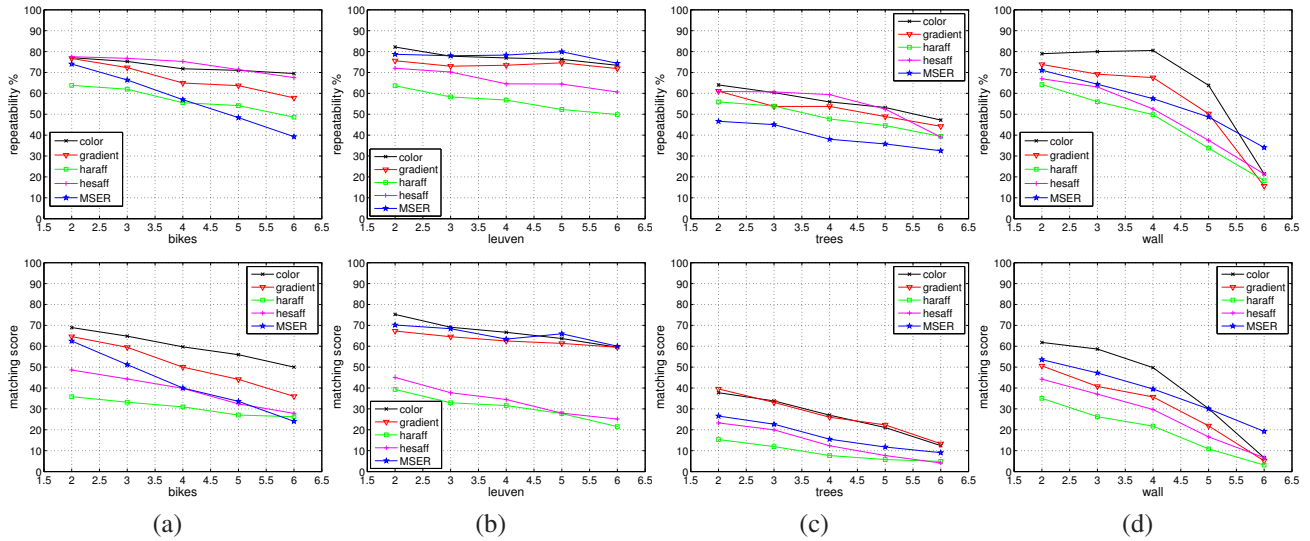

Figure 3. Repeatability and matching score. (a) Blur for the structured scene. (b) Illumination change. (c) Blur for the textured scene. (d) Viewpoint change for the textured scene. Our detectors perform very well on these data sets, in comparison with the best results reported in [21].

4.2. Action Classification with Histogram-based Interest Point Detectors in Space-Time

In this experiment, we show how to extend our histogram-based method to the detection of space-time interest points, and apply the detector to the problem of action classification. Similar ideas have been explored in [14], [25], [26] for event and action analysis.

Detecting Space-Time Interest Points. In this section we show how to use our method to detect space-time interest points in a video. We employ the human action database provided by [10]. The database contains ten kinds of actions, and for each action we use nine videos of the same action performed by different people. We ignore the silhouette information since our method does not assume a known background. We choose the 64-bin oriented gradient histogram to perform our detector, and use a 3D Gaussian with $\sigma = 2$ for computing the weighted histogram. Given a video
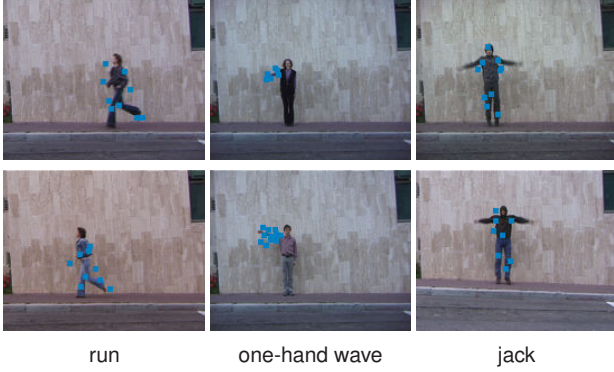
run one-hand wave jack

Figure 4. Examples of detected space-time interest points for three types of actions.

	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10
a1	88.9	0	0	0	11.1	0	0	0	0	0
a2	0	88.9	11.1	0	0	0	0	0	0	0
a3	0	33.3	44.4	0	22.2	0	0	0	0	0
a4	0	0	0	100	0	0	0	0	0	0
a5	0	0	22.2	0	77.8	0	0	0	0	0
a6	0	0	0	11.1	0	88.9	0	0	0	0
a7	0	0	0	0	0	0	100	0	0	0
a8	0	0	0	11.1	0	0	0	66.7	22.2	0
a9	0	0	0	0	0	0	0	11.1	88.9	0
a10	0	0	0	0	0	0	0	0	0	100

Table 1. Action confusion. a1-walk, a2-run, a3-skip, a4-jack, a5-jump, a6-pjump, a7-side, a8-wave1, a9-wave2, and a10-bend.

sequence, we compute the 3-by-3 Hessian matrix $H(x, y, t)$ of a *space-time cube* surrounding each pixel in each frame. The 3-by-3 Hessian matrix can be easily derived from Eqs. (4)-(9) by including the terms related to the time domain. In our experiment the size of a space-time cube is $15 \times 15 \times 9$ pixels, *i.e.*, a cube consists of nine 15-by-15 patches extracted from the nine consecutive frames. The space-time locations (x, y, t) with locally maximal responses of $R(H)$ are selected as candidates. We sort the candidates by the response values and keep the top 270 candidates to form the set of space-time interest points. Some examples of detection results are shown in Fig. 4.

Action Classification. We try to make use of the space-time cubes extracted by our histogram-based detector to solve the problem of action classification on videos. Given a video sequence S_i, we carry out the aforementioned scheme to select 270 interest points. The space-time cube corresponding to each interest point generates a 1152-dimensional feature vector by stacking the 128-dimensional SIFT descriptions in nine frames. We then use affinity propagation [9] to group the feature vectors in sequence S_i into K_i clusters. To compare two sequences S_i and S_j, we combine the cluster centers of S_i and S_j to get $(K_i + K_j)$ centers, and reassign the feature vectors in each sequence to the new set of $(K_i + K_j)$ centers. As a result, we obtain two $(K_i + K_j)$-bin histograms for the two sequences. The similarity between S_i and S_j is given by the Bhattacharyya coefficient between the two histograms. For every video sequence, we perform a leave-one-out test procedure. On each test, we remove the test sequence from the database and use the nearest-neighbor method to determine the class of the test sequence. That is, for test sequence S_i, we find the most similar sequence S_{j*} in the database according to the Bhattacharyya similarity measure. The result of leave-one-out test is shown in Table 1. The average precision is 84.4%. Our result is comparable to the state-of-the-art (82.6% by

[25]) that draws on more complex 3D SIFT descriptors and uses SVM to train classifiers. In [25], the reported average precision of using stacking 2D SIFT descriptions is only 47.8%. Note that the precision of leave-one-out test on this database can be further improved if the background and silhouette information is used to analyze the space-time shapes [10], which is beyond the scope of this paper.

5. Conclusion

This paper presents a new criterion for detecting interest points. The integration of low-level feature histograms and interest-point detectors is achieved by introducing the histogram-based similarity measure into the analysis of Hessian matrix. Since most of the popular local region descriptors are derived from histograms of low-level features rather than pixel-wise representations, our method equips local feature detectors with similar representations to the descriptors. The experimental results show that the histogram-based interest point detectors are less sensitive to small-scale variations and thus more effective in matching textured scenes under blur and illumination changes. In addition to the two histogram-based representations presented in this paper, other types of low-level features such as Gabor filter responses can also be incorporated into our framework for constructing histograms. The promising results of solving matching and action classification problems suggest that the histogram-based detectors should be useful for more applications in computer vision.

Appendix

The Bhattacharyya coefficient ρ between two distributions $p = \{p_k\}_{k=1,...,L}$ and $q = \{q_k\}_{k=1,...,L}$ is defined by $\rho = \sum_{k=1}^{L} \sqrt{p_k \, q_k}$. Suppose that q is the target distribution, and $p(\mathbf{y})$ is a spatial-varying distribution depending on location $\mathbf{y}$. If we apply a shift $\delta\mathbf{y}$ to $\mathbf{y}$, the Bhattacharyya coefficient between

$p(\mathbf{y} + \delta\mathbf{y})$ and q can be approximated by the second-order Taylor expansion

$$\rho \quad = \sum_{k=1}^{L} \sqrt{p_k(\mathbf{y} + \delta\mathbf{y}) \cdot q_k} \tag{12}$$

$$\approx \quad \sum_{k=1}^{L} \sqrt{p_k(\mathbf{y}) q_k} + \frac{1}{2} \sum_{k=1}^{L} \nabla p_k(\mathbf{y}) \delta\mathbf{y} \sqrt{\frac{q_k}{p_k(\mathbf{y})}}$$

$$+ \frac{1}{2} \delta\mathbf{y}^T H(\mathbf{y}) \, \delta\mathbf{y} \tag{13}$$

where $H(\mathbf{y})$ is the Hessian matrix of ρ at the location $\mathbf{y}$. We may replace $\nabla p_k(\mathbf{y})\delta\mathbf{y}$ by $(p_k(\mathbf{y} + \delta\mathbf{y}) - p_k(\mathbf{y}))$ and obtain

$$\rho \quad = \sum_{k=1}^{L} \sqrt{p_k(\mathbf{y} + \delta\mathbf{y}) \cdot q_k}$$

$$\approx \quad \frac{1}{2} \sum_{k=1}^{L} \sqrt{p_k(\mathbf{y}) q_k} + \frac{1}{2} \sum_{k=1}^{L} p_k(\mathbf{y} + \delta\mathbf{y}) \sqrt{\frac{q_k}{p_k(\mathbf{y})}}$$

$$+ \frac{1}{2} \delta\mathbf{y}^T H(\mathbf{y}) \, \delta\mathbf{y} \,. \tag{14}$$

Let $q = p(\mathbf{y})$, and the auto-similarity based on the Bhattacharyya coefficient is given by

$$\tilde{\rho} \quad = \sum_{k=1}^{L} \sqrt{p_k(\mathbf{y} + \delta\mathbf{y}) \cdot p_k(\mathbf{y})}$$

$$\approx \quad \frac{1}{2} \sum_{k=1}^{L} \sqrt{p_k(\mathbf{y}) p_k(\mathbf{y})} + \frac{1}{2} \sum_{k=1}^{L} p_k(\mathbf{y} + \delta\mathbf{y}) \sqrt{\frac{p_k(\mathbf{y})}{p_k(\mathbf{y})}}$$

$$+ \frac{1}{2} \delta\mathbf{y}^T H(\mathbf{y}) \, \delta\mathbf{y} \,. \tag{15}$$

Since $p(\mathbf{y})$ is a probability distribution, we have $\sum_{k=1}^{L} p_k(\mathbf{y}) = \sum_{k=1}^{L} p_k(\mathbf{y} + \delta\mathbf{y}) = 1$, and it follows that

$$\tilde{\rho} \approx 1 + \frac{1}{2} \delta\mathbf{y}^T H(\mathbf{y}) \, \delta\mathbf{y} \,. \tag{16}$$

References

[1] H. Bay, T. Tuytelaars, and L. J. V. Gool. Surf: Speeded up robust features. In *ECCV (1)*, pages 404–417, 2006.

[2] P. Beaudet. Rotationally invariant image operators. In *4th Int. Joint Conf. Patt. Recog.*, pages 579–583, 1978.

[3] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–110, 1943.

[4] M. Brown and D. G. Lowe. Recognising panoramas. In *ICCV*, pages 1218–1227, 2003.

[5] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *CVPR (2)*, pages 142–149, 2000.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR (1)*, pages 886–893, 2005.

[7] G. Dorkó and C. Schmid. Maximally stable local description for scale selection. In *ECCV (4)*, pages 504–516, 2006.

[8] R. Fergus, F.-F. Li, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *ICCV*, pages 1816–1823, 2005.

[9] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

[10] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(12):2247–2253, 2007.

[11] C. Harris and M. Stephens. A combined corner and edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 147–151, 1988.

[12] T. Kadir and M. Brady. Saliency, scale and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.

[13] T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. In *ECCV (1)*, pages 228–241, 2004.

[14] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, pages 432–439, 2003.

[15] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.

[16] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, 1999.

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[18] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *ECCV (1)*, pages 128–142, 2002.

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

[20] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10):1615–1630, 2005.

[21] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. J. V. Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1-2):43–72, 2005.

[22] P. Montesinos, V. Gouet, R. Deriche, and D. Pelé. Matching color uncalibrated images using differential invariants. *Image Vision Comput.*, 18(9):659–671, 2000.

[23] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(5):530–535, 1997.

[24] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172, 2000.

[25] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *ACM Multimedia*, pages 357–360, 2007.

[26] E. Shechtman and M. Irani. Space-time behavior based correlation. In *CVPR (1)*, pages 405–412, 2005.

[27] J. Sivic, F. Schaffalitzky, and A. Zisserman. Object level grouping for video shots. In *ECCV (2)*, pages 85–98, 2004.

[28] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: A survey. *Foundations and Trends in Computer Graphics and Vision*, 3(3):177–280, 2007.