Spatiotemporal Phase Congruency Based Invariant Features for Human Behavior Classification

Hao Wang[1,2], Yi Xu[1,2], Xiaokang Yang[1,2], Li Song[1,2], and Wenjun Zhang[1,2]

[1] Institute of Image Communication and Information Processing
[2] Shanghai Key Laboratory of Digital Media Processing and Transmissions
{wang.hao.sjtu2010,xuyi,xkyang,song_li,
zhangwenjun}@sjtu.edu.cn

Abstract. In applications of behavior recognition, the use of spatiotemporal invariant feature points can improve the robustness to noise, illumination and geometric distortions. In this paper, we develop a novel detection model of spatiotemporal invariant feature by generalizing the notion of image phase congruency to video volume phase congruency. The proposed model detects feature points by measuring the spatiotemporal phase congruency of Fourier series components along with their characteristic scale and principal orientation. Compared with other state-of-the-art methods, the key advantages of this interest point detector include the invariance to contrast variations and more precise feature location. Furthermore, an invariant feature descriptor is advanced based on the phase congruency map, resulting in enhanced discriminative power in classification tasks. Experimental results on KTH human motion dataset demonstrate the validity and effectiveness of the extracted invariant features in the human behavior recognition scheme.

Keywords: Spatiotemporal phase congruency, invariant feature, human behavior recognition.

1 Introduction

Detecting, recognizing and classifying human actions in video sequences have received a lot of attention in recent years. However, there exist significant barriers in these tasks because the objects in video can vary in posture, appearance and size, in addition to camera motion changes, view-point changes and occlusions.

These challenges can be partly overcome by using spatiotemporal feature point. Spatiotemporal feature point, according to Dollar [1], is a short, local video sequence containing rich information of object action. Such spatiotemporal interest points are able to locate an action and contain important information of both object appearance and velocity. Even though many feature point detection algorithms are available for image analysis, much less work has been done in video domain. Laptev [3] extended the Harris corner detection algorithm [4] to spatiotemporal domain by incorporating temporal derivatives into second-moment matrix. Another method proposed by Dollar [1] is dependent on separable space-time linear filters to extract sharp variations in

P. Muneesawang et al. (Eds.): PCM 2009, LNCS 5879, pp. 100–109, 2009.

video content. Oikonomopoulos [5] generalized the idea of saliency regions in spatial images to the spatiotemporal case, based on the work of Kadir and Brady [6]. However, the common problems of these methods are that their responses vary considerably with image contrast and the use of Gaussian smoothing also makes the feature location less accurate.

In this paper, we will introduce a novel detection model of spatiotemporal invariant features by generalizing the notion of 2D image phase congruency to 3D volumetric phase congruency. Phase congruency algorithm was proposed by Kovesi [7] [8] [9] to capture spatial interest points. Following work of Myerscough adopted it directly to compute temporal phase congruency [16]. Here we generalize the concept of phase congruency in video volume analysis, not simply regarding it as the direct 3D counterpart of the original model. A cost-efficient way of orientation selection is suggested in spatiotemporal filter design. Characteristic scale and principal orientation of each spatiotemporal feature point are computed to determine its local volume support. An invariant feature descriptor is developed using spatiotemporal phase congruency map rather than the raw data of video volume. The key advantages of these spatiotemporal interest points include the invariance to contrast variations, precise feature location and enhanced discriminative power in classification tasks.

The rest content of this paper is structured as follows: In section 2, we review the image phase congruency algorithm and present our spatiotemporal phase congruency model for interest point detection. This is followed by section 3 covering the issues involved in establishing invariant spatiotemporal features. In section 4, comparison experiments are presented to evaluate the discriminative power of these proposed invariant features in human behavior classification task. Finally, conclusion remark is drawn in section 5.

2 Spatiotemporal Phase Congruency Feature Detection

2.1 Image Phase Congruency Model

Image phase congruency model is developed by Kovesi [8]. The importance of phase information for image representation is well demonstrated by Oppenheim and Lim [10]. Also a great variety of feature types are found to give rise to high phase congruency, such as step edges, line and roof edges and Mach bands [11].

Venkatesh and Owens [13] proposed to calculate and search for peaks of local energy function to identify points of maximum phase congruency. The local energy function of signal $I(x)$ is defined as

$$E(x) = \sqrt{F^2(x) + H^2(x)} \tag{1}$$

where $F(x)$ is the signal $I(x)$ with its DC component removed and $H(x)$ is its Hilbert counterpart. The components $F(x)$ and $H(x)$ are obtained by convolving signal $I(x)$ with quadrature pair of filters. Venkatesh and Owens formulated the relationship between local energy function and phase congruency as,

$$E(x) = PC(x)\sum_{n} A_n(x) \tag{2}$$

where A_n represents the *nth* Fourier series component of signal $I(x)$.

Kovesi [8] proposed to calculate phase congruency using log Gabor filter, which is a band pass filter without sensitivity to DC component. The transfer function of log Gabor filter is,

$$g(w) = e^{\frac{-(log\ (w/w_0))^2}{2(log\ (k/w_0))^2}} \tag{3}$$

where w_0 is the center frequency, parameter k is a constant for determining the number of octave bands. In order to take all the Fourier series components into account, we need to design a series of log Gabor filters with different center frequencies so that altogether they are able to uniformly cover the entire frequency domain.

In (2), the cosine of the phase deviation is implicitly exploited to measure phase congruency. Although it is theoretically correct, in practice this measure is not sensitive enough to phase deviations. So Kovesi [8] developed a more sensitive measure for phase deviation to replace simple cosine function as

$$\Delta\Phi(x) = cos(\emptyset_n(x) - \overline{\emptyset}(x)) - \left|\sin(\emptyset_n(x) - \overline{\emptyset}(x))\right| \tag{4}$$

where $\emptyset_n(x)$ is the phase angle of the *nth* Fourier component and $\overline{\emptyset}(x)$ is the weighted mean phase angle of all Fourier components. Then the new measure of phase congruency becomes,

$$PC(x) = \frac{\sum_n W(x)\lfloor A_n(x)\Delta\Phi(x) - T\rfloor}{\sum_n A_n(x)} \tag{5}$$

where $W(x)$ is the frequency spread weighting function which penalizes points with a narrow frequency spread. Noise threshold T is estimated based on the response of the filter of the smallest scale. The notation $\lfloor x\rfloor$ equals to $max\ (x, 0)$.

In order to analyze two-dimensional signals, such as images, we have to apply the one-dimensional analysis over a series of different orientations and then combine the results. Kovesi [8] suggested to extend 1D log-Gabor filters defined in (3) to 2D case by adding an angular spreading function,

$$G(\theta) = e^{-\frac{(\theta-\theta_0)}{2\sigma_\theta^2}} \tag{6}$$

where θ_0 is the orientation of the filter and σ_θ is the standard deviation of the Gaussian spreading function in the angular direction. After combining the result over o different orientations, the phase congruency measure becomes

$$PC(x) = \frac{\sum_o \sum_n W(x)\lfloor A_n(x)\Delta\Phi(x) - T\rfloor}{\sum_n A_n(x)} \tag{7}$$

2.2 Spatiotemporal Phase Congruency Model for Detecting Interest Points

To establish a useful analytical tool for video sequences, the generalization of phase congruency model for spatiotemporal interest point detection is realized in a more complex way than the direct 3D counterpart of image phase congruency.

In Figure 1, we illustrate the workflow of interest point extraction based on spatiotemporal phase congruency. First phase congruency values are calculated in N orientations in space-time domain, and then combined to obtain a spatiotemporal phase congruency map. After non-maximal suppression and greedy clustering, spatiotemporal interest points are finally extracted along with their characteristic spatial and temporal scales.

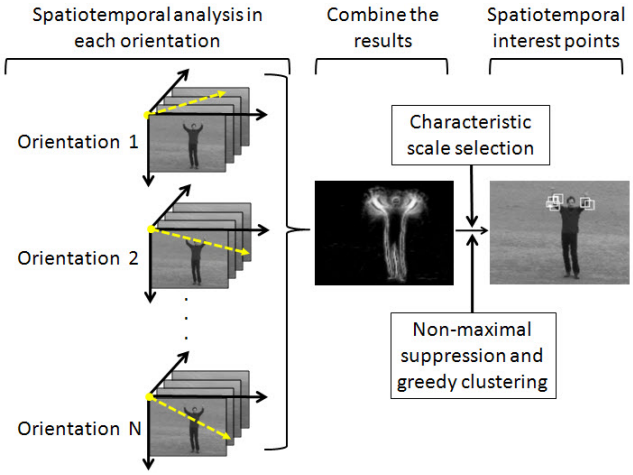

Fig. 1. Workflow of interest point extraction based on spatiotemporal phase congruency

2.2.1 Spatiotemporal Log Gabor Filters with Octave Bands

The centerpiece of phase congruency algorithm is the filter design. As we described above, the filter can be separated into radial component as (3) and angular component as (6). As for spatiotemporal log Gabor filters, the radial component is designed in the same way as formulated in (3), while the angular component requires modification.

It needs to first define the orientation of the angular component. Unlike in spatial domain where we need only one angle to specify an orientation, in spatiotemporal domain we need two. Suppose that *XYT* space provides spatiotemporal representation of video volume, the *UVW* space is the corresponding Fourier frequency space. As in Figure 2, we can define an arbitrary filter orientation vector $\overrightarrow{orient}$ by specifying two angles α_W and α_{UV} as follows,

$$\overrightarrow{orient}(\alpha_W, \alpha_{UV}) = (\cos \alpha_{UV}, \sin \alpha_{UV}, \cot \alpha_W) \tag{8}$$

Given filter orientation $\overrightarrow{orient}$, we can calculate the angle $\theta(\overrightarrow{orient}, \vec{v})$ between it and an arbitrary UVW space vector $\vec{v} = (u, v, w)$, that is

$$\theta(\overrightarrow{orient}, \vec{v}) = \cos^{-1}\left(\frac{\overrightarrow{orient} \cdot \vec{v}}{|\overrightarrow{orient}||\vec{v}|}\right) \tag{9}$$

Thus the angular component of the spatiotemporal log Gabor filter is constructed as

$$G(\theta) = e^{-\frac{\theta(\overrightarrow{orient}, \vec{v})}{2\sigma_\theta^2}} \tag{10}$$

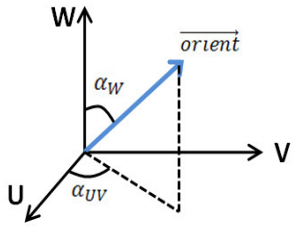

Fig. 2. An arbitrary orientation vector $\overrightarrow{orient}$ specified in UVW space by two angles α_W and α_{UV}. α_{UV} is the angle between U axis and the projection of orientation $\overrightarrow{orient}$ on UV plane. α_W is the angle between orientation $\overrightarrow{orient}$ and W axis.

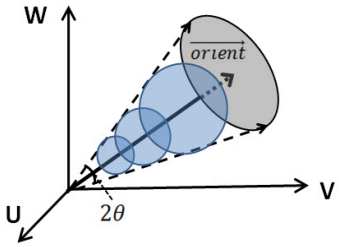

Fig. 3. Illustration of orientational spatiotemporal log Gabor filters

In Figure 3, we illustrate the constructed spatiotemporal filters, the supports of which are circled by blue color. The cone region is specified by the filter's angular component and the center frequencies along the orientation vector are specified by the radial component. The aperture of 2θ is controlled by parameter σ_θ in formula (10). Three octave-band filters are constructed here to ensure the spectral coverage along this orientation.

2.2.2 Computation of Spatiotemporal Phase Congruency

With the constructed spatiotemporal log Gabor filters, we are able to calculate phase congruency in any 3D orientation. Kovesi has shown in image phase congruency algorithm that orientation interval of 30 degrees provides good result. Accordingly, we select 30 degrees interval for quantizing α_W and α_{UV}(0°, 30°, 60°, 90°, 120°, 150° for each). However it should be noted that spatiotemporal feature points must be prominent in both spatial domain and temporal domain. For orientations with α_W of 90°, the algorithm can only calculate the phase congruency in spatial domain. On the contrary, we are only looking at the temporal domain if the orientation component α_W equals 0 degrees. Whenever the phase congruencies produced by these orientations ($\alpha_W = 0° \; or \; 90\,°$) are too high or too low, the resulting feature points will be biased. We propose to calculate only in the directions where both spatial and temporal information are taken into account. Hence, α_W should only take values of 30, 60, 120 and 150 degrees. Thus, we both enhance the effectiveness of spatiotemporal feature and improve the cost efficiency by removing unnecessary orientations.

The spatiotemporal phase congruency model is established by combining results in different orientations specified by α_W and α_{UV}

$$PC(x) = \frac{\sum_{\alpha_W} \sum_{\alpha_{UV}} \sum_n W(x) \lfloor A_n(x) \Delta\Phi(x) - T \rfloor}{\sum_n A_n\,(x)} \tag{11}$$

where the noise threshold T is calculated in the same fashion as in the image phase congruency model proposed by Kovesi. Similar to image phase congruency, spatiotemporal phase congruency is also a dimensionless quantity invariant to contrast. In absence of Gaussian smoothing, spatiotemporal phase congruency can locate the interest points in video volume more precisely.

3 Invariant Feature Description of Spatiotemporal Interest Points

Robust spatiotemporal feature detection is reliable with regard to the significance of phase congruency values. In this section, we deliberate on another important issue for subsequent human behavior classification task, that is, invariant feature description.

3.1 Spatiotemporal Characteristic Scale for Feature Description

To achieve scale-invariant descriptions of spatiotemporal interest points, we need to specify an adaptive support around the feature point according to its characteristic scales, which is called a cuboid by Dollar [1]. To specify a cuboid, we need not only the central location of the feature point, but also the spatial scale σ_s and the temporal scale σ_t of the neighborhood region.

The characteristic scale of the feature points are defined as the scale corresponding to the filter support that produces the strongest response. Assuming the filter produces the strongest response around a spatiotemporal interest point at characteristic scale s and principal orientation component α_W, the temporal characteristic scale σ_t is then calculated as the projection of scale s on W (T) axis and the spatial characteristic scale σ_s is calculated as the projection of scale s on UV plane (XY plane).

$$\sigma_t = \mathrm{s} \times \cos \alpha_{\mathrm{W}} \tag{12}$$

$$\sigma_s = \mathrm{s} \times \sin \alpha_{\mathrm{W}} \tag{13}$$

So far we can obtain the phase congruency value and spatiotemporal characteristic scales of each point in video sequences. Then, thresholding and non-maximal suppression are applied to eliminate the points with non-maximal phase congruency values along their principal orientations. This is followed by greedy clustering algorithm to form spatiotemporal regions through grouping spatiotemporal points with similar location and scale. By doing so, we eliminate the redundancy of the neighboring feature points and maintain the discriminative power of the underlying features.

3.2 Invariant Feature Descriptor Based on Phase Congruency Map

In conventional methods, the raw grayscale video data is used to establish the feature descriptor. As shown in Figure 4, the phase congruency map exhibits strong advantages over raw video data. It not only highlights the edges and corners with noise removed, but also filters out the quasi-static background, leaving only the dynamic objects that interest us. Since phase congruency map better represents the structure and motion of the video content, we exploit it to generate superior feature descriptors. Since phase congruency map is generated in the feature detection step, it does not introduce any additional computational overhead.

'Bag of words' model [15] is employed to execute the task. According to 'bag of words' model, the codebook is first formed. Then all the invariant spatiotemporal features are matched to the codebook. As a result, the video sequence is represented by the frequency distribution of the code words, which would be used as the input of the classifier.

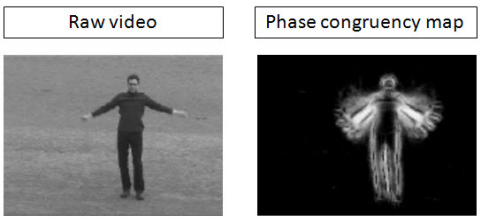

Fig. 4. Comparison of raw video data and phase congruency map

4 Experimental Results

4.1 Spatiotemporal Interest Point Detection

We applied our algorithm to detect spatiotemporal interest points in KTH human motion dataset (Schuldt [14]). The feature detection results of our approach are compared with those from other state-of-the-art interest point detectors, including the separable filter method proposed by Dollar [1] and the Harris corner point method proposed by Laptev [3]. As for the two latter methods, the default parameter setting is used. The comparison results are listed in Figure 5. Obviously, our approach demonstrates the strength of superior detection performance, producing a relatively large amount of feature points with higher precision.

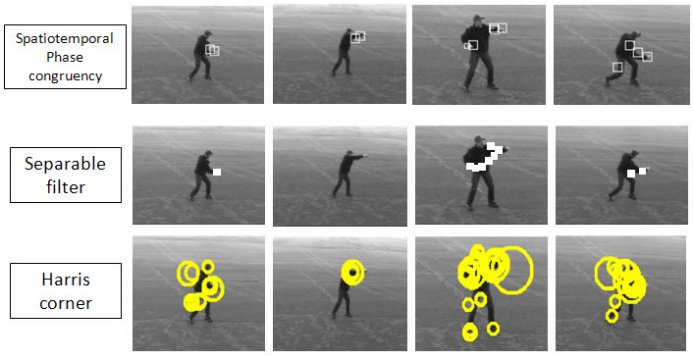

Fig. 5. Comparison of three spatiotemporal feature detectors

4.2 Human Behavior Classification

In the experiment, we use the Support Vector Machine (SVM) classifier and KTH human motion dataset. The codebook size is set at 1000. Three fourths of the video clips are used for training and one fourth for testing. The results we show below are verified to be relatively stable within a range of different parameter settings.

First, we compare the performance of two invariant feature descriptors, which are respectively generated by raw video data and phase congruency map. It is shown in Figure 6 that the result using phase congruency based descriptor outperforms the other one in terms of recognition accuracy.

Furthermore, we show in Figure 7 the confusion matrix for KTH dataset using phase congruency value as descriptor input, where the rows are true categories and the columns are the classification results. Higher recognition accuracy (86.7%) is achieved in our approach in comparison with the result of Dollar [1] (81.2%) and Schuldt [14] (71.7%).

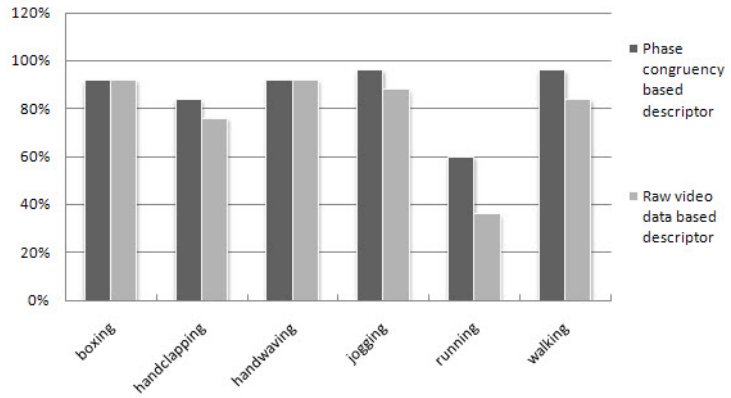

Fig. 6. Comparison of two invariant feature descriptors with regard to the classification accuracy

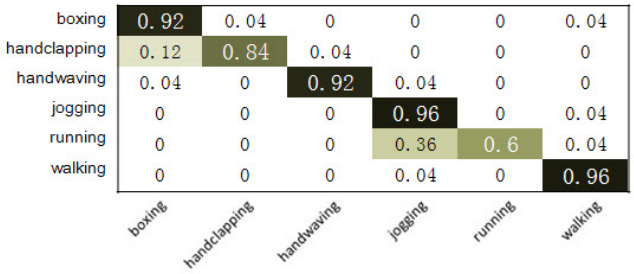

Fig. 7. Confusion matrix of our classification result

5 Conclusion

In this paper, a novel spatiotemporal feature detector is proposed by generalizing the notion of image phase congruency to spatiotemporal domain. A superior feature descriptor is then advanced using spatiotemporal phase congruency map, which demonstrates its strength in human behavior classification task.

In future work, we would introduce the proposed approach to other motion recognition related applications. In addition, recognition accuracy might be further enhanced by investigating into the classification scheme.

Acknowledgments. This work was supported in part by Research Fund for the Doctoral Program of Higher Education of China (200802481006), NCET-06-0409, Cultivation Fund of the Key Scientific and Technical Innovation Project of MOE (706022), and the 111 Project (B07022).

References

1. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior Recognition via Sparse Spatio-Temporal Features. In: ICCV VS-PETS 2005, Beijing, China (2005)
2. Haralick, R., Shapiro, L.: Computer and Robot Vision II. Addison-Wesley, Reading (1993)
3. Laptev, I.: On Space-Time Interest Points. International Journal of Computer Vision 64(2/3), 107–123 (2005)
4. Harris, C., Stephens: A combined corner and edge detector. In: Alvey Vision Conference, pp. 147–152 (1988)
5. Oikonomopoulos, A., Patras, I., Pantic, M.: Spatiotemporal saliency for human action recognition. In: ICME 2005, pp. 430–433 (2005)
6. Kadir, T., Brady, M.: Scale saliency: a novel approach to salient feature and scale selection. In: International Conference on Visual Information Engineering, November 2000, pp. 25–28 (2000)
7. Kovesi, P.D.: A Dimensionless Measure of Edge Significance from Phase Congruency Calculated via Wavelets. In: The First New Zealand Conference on Image and Vision Computing, Auckland, August 16-18, pp. 87–94 (1993)
8. Kovesi, P.: Image Features From Phase Congruency. Videre: A Journal of Computer Vision Research 1(3) (Summer 1999)
9. Kovesi, P.: Phase Congruency Detects Corners and Edges. In: The Australian Pattern Recognition Society Conference: DICTA 2003, Sydney, December 2003, pp. 309–318 (2003)
10. Oppenheim, A.V., Lim, J.S.: The importance of phase in signal. Proceedings of the IEEE 69, 529–541 (1981)
11. Morrone, M.C., Ross, J.R., Burr, D.C., Owens, R.A.: Mach bands are phase dependent. Nature 324(6094), 250–253 (1986)
12. Morrone, M.C., Owens, R.A.: Feature detection from local energy. Pattern Recognition Letters 6, 303–313 (1987)
13. Venkatesh, S., Owens, R.A.: An energy feature detection scheme. In: The International Conference on Image Processing, Singapore, pp. 553–557 (1989)
14. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: ICPR, pp. 32–36 (2004)
15. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)
16. Myerscough, P.J., Nixon, M.S.: Temporal phase congruency. In: The 6th. IEEE Southwest Symposium on Image Analysis and Interpretation, pp. 76–79 (2004)