SURF- and Optical Flow-based Action Recognition with Outlier Management

Md. Atiqur Rahman Ahad
Kyushu Institute of Technology
1-1, Sensui, Tobata, Kitakyushu, Japan

J. Tan, H. Kim, S. Ishikawa
Kyushu Institute of Technology, Japan
{atiqahad,etheltan,ishikawa}@ss10.cntl.kyutech.ac.jp

Abstract

A spatio-temporal method is developed to recognize various activities by considering local interest points to compute global features. The key interest points are computed by Speeded-Up Robust Features (SURF). With the key points, we employ gradient-based optical flow, exploit RANSAC to remove outliers, and then split flow vectors into different channels. Based on these robust flow vectors, we compute history and energy templates to represent each activity. However, presence of various unwanted corner points in outdoor scenes may deter better representation for an action, mainly in cluttered environment. So, frame-subtracted accumulated image are exploited to mask out unwanted points for robust action-representations. These are employed to recognize actions.

1. Introduction

Human action understanding have various applications in computer vision [1-2]. Here, we extend our spatio-temporal template-based and appearance-based method [3] – to represent and understand some complex actions and interactions in outdoor scene with cluttered environment. We consider the Motion History Image (MHI) [1,4] as base of our development. One of the key constraints of this method is that it cannot solve the motion self-occlusion problem due to motion overriding [1,5]. We ponder to solve this issue so that complex actions as well as multiple persons' interactions can be addressed reasonably well. It considers local interest points to compute global features for various action representations. The key interest points are computed by Speeded-Up Robust Features (SURF), which is a scale- and rotation-invariant interest point detector and descriptor [6]. With the key points, we employ optical flow and apply RANSAC (RANdom Sample Consensus) [7] to reduce outliers. Then we split the gradient-based optical flow into different channels. Afterwards, based on four different flow vectors, we compute motion history and energy images to represent each activity. Finally, frame-subtracted accumulated image is masked to remove unwanted corner points in the scene. Due to the nature of the flow vectors, the interactions can be isolated in left, right, up and down directions. The proposed SURF-based History Image and Energy Image with outlier management demonstrate better image representations due to its SURF-based local key interest points and later to the corresponding gradient-based optical flow-vectors from the interest points. Fig. 1 demonstrates a basic flow diagram of this approach.

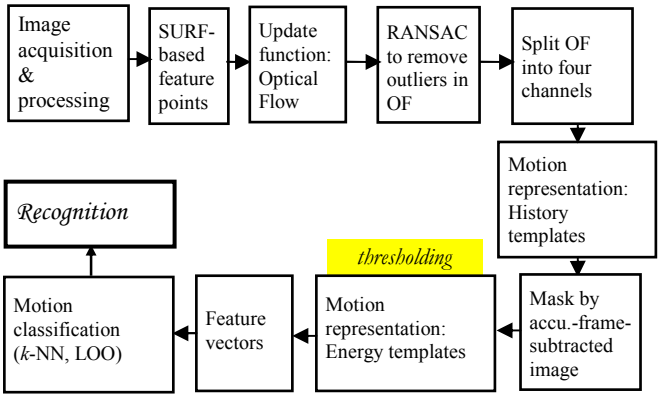


Figure 1. Basic flow diagram of the proposed method for recognition.

2. Development of the Method

In SURF, the determinant of the Hessian is employed. The Hessian matrix is roughly approximated by using a set of box-type filters. Given a point $\mathbf{x} = (x, y)$ in an image I, the Hessian matrix $H(\mathbf{x},\sigma)$ in $\mathbf{x}$ at scale σ is defined as, $H(\mathbf{x},\sigma) = \begin{bmatrix} L_{xx}(\mathbf{x},\sigma) & L_{xy}(\mathbf{x},\sigma) \\ L_{xy}(\mathbf{x},\sigma) & L_{yy}(\mathbf{x},\sigma) \end{bmatrix}$. Here $L_{xx}(\mathbf{x},\sigma)$ is the convolution of the Gaussian second order derivative $\frac{\partial^2}{\partial x^2} g(\sigma)$ with the image I in point $\mathbf{x}$ and similarly for $L_{xy}(\mathbf{x},\sigma)$ and $L_{yy}(\mathbf{x},\sigma)$. These Gaussians derivatives are approximated (e.g., $D_{xx}(\mathbf{x},\sigma)$, $D_{xy}(\mathbf{x},\sigma)$ and $D_{yy}(\mathbf{x},\sigma)$) by considering approximated box-type filters. The approximated determinant of the hessian matrix represents the blob response in the image at location $\mathbf{x}$. Then these responses are kept in a blob response map. So we get, $\det(H_{\text{approx}}) = D_{xx}D_{yy} - (0.9D_{xy})^2$. We consider the key interest points from consecutive frames and then compute the optical flow based on those local feature-points. Optical flow-based update function, $\Psi(x, y, t)$ is computed directly from consecutive frames. Then we employ RANSAC to remove outliers of the flow vectors [7]. These are split into four different channels ($\varpi \in \{left, right, up, down\}$) to construct four-directional motion templates based on the

SURF-extracted interest points-based flow vectors. For each direction, based on a threshold ξ on pixel value, the templates are computed as,

$$\wp SbHI_{\tau}^{\varpi}(x,y,t) = \begin{cases} \tau & \text{if } \Psi^{\varpi}(x,y,t) > \xi \\ \max(0, \wp SbHI_{\tau}^{\varpi}(x,y,t-1) - \delta) & \text{otherwise} \end{cases}$$

One key concern in outdoor scene or in cluttered background is the presence of unwanted corner points due to corner point detectors. We mask out these unnecessary points intuitively by accumulating frame-subtracted image, as it consumes the entire motion information regions and hence, masking with it, we can clean-up the unnecessary corner points and holes in the final representations for an action. After having $\wp SbHI_{\tau}(x,y,t)$, we compute the energy images by thresholding it over zero.

3. Experimental Results and Discussion

After having the image representations from action video, feature vectors are extracted from each action by exploiting the seven moment invariants by Hu for each template [8]. For comparison, we employed directional motion history image ($DMHI_{\tau}$) [5], frame-subtraction-based basic motion history image [4] and optical flow-based history image (OHI_{τ}) method (proposed here). We employ nearest neighbor algorithm and leave-one-out scheme. We employ the methods in a challenging outdoor database of different actions with a background that is cluttered and in changing illumination from an uncalibrated frontal-view camera. Fig. 2 illustrates some sequential frames for each action of the outdoor dataset (9 actions by eight subjects).

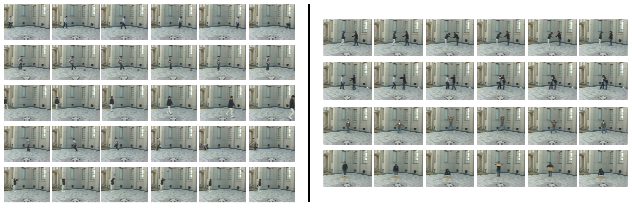


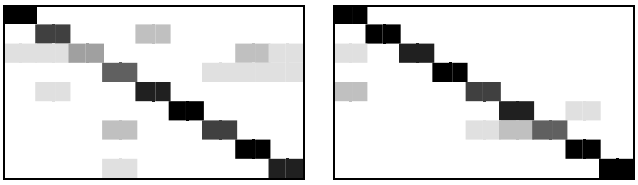
Figure 2. Outdoor action dataset: each row (of each block) represents one action – from A1~A9 (walk; kick a ball; diagonal walk; hopscotch; ball-throwing; handshake; hug; jumping jack; & lifting box from floor.).

TABLE I. Comparative recognition results

Action#	OHI_{τ}	MHI_{τ}	$DMHI_{\tau}$	$\wp SbHI_{\tau}$
A1	75.0	100	62.5	100
A2	37.5	75.0	62.5	100
A3	25.0	37.5	50.0	87.5
A4	62.5	62.5	87.5	100
A5	75.0	87.5	87.5	75.0
A6	50.0	100	100	87.5
A7	50.0	75.0	75.0	62.5
A8	62.5	100	100	100
A9	100	87.5	100	100
Average	**59.7**	**80.5**	**83.3**	**90.3**

Usually, recognition in outdoor environment is challenging due to the various reflections, changing in illumination, cluttered environment, and edges. But using our proposed method ($\wp SbHI_{\tau}(x,y,t)$), more than 90.3% overall recognition result have been achieved (Table I). As seen from the Table I, the average recognition results for the basic MHI method is 80.5%; the OHI_{τ} method is 59.7%; and the directional motion history method is 83.3%. Five of the actions produced 100% recognition rate with our method. As evident from the confusion matrix of $\wp SbHI_{\tau}(x,y,t)$ (of Table II), it is noticeable that A7 confuses with three other actions and A5 mimics with A1 in two different cases (Table II(b)). Therefore, the impacts of A7 and A5 reduce the average recognition rate.

TABLE II. Confusion matrices for (a) MHI & (b) $\wp SbHI_{\tau}$



(a) Confusion matrix of MHI_{τ} (b) Confusion matrix of $\wp SbHI_{\tau}$

4. Conclusions

We explore a spatio-temporal template-based method to represent and understand different actions in complex outdoor environment. The method is based on the SURF-based local feature point detection. The inclusion of the accumulated-frame-subtracted image to eliminate non-moving contents and outliers after RANSAC, crafts the method a robust one. In future, we will work on tuning the method for better performance in more complex actions and multiple persons' interactions. The presence of camera motion can make this method a bit difficult to produce sound result. It is important to explore that area too.

5. References

[1] M. Ahad, J. Tan, H. Kim, and S. Ishikawa, "Motion history image: its variants and applications", Machine Vision and Applications, 1-27, 2010.

[2] M. Ahad, J. Tan, H. Kim, and S. Ishikawa, "Human activity recognition: various paradigms", Proc. Int. Conf. on Control, Automation and Systems, 1896-1901, 2008.

[3] M. Ahad, J. Tan, H. Kim, and S. Ishikawa, "SURF-based spatio-temporal history image method for action representation", Int. Conf. on Industrial Technology, 411-416, 2011.

[4] A. Bobick and J. Davis, "The recognition of human movement using temporal templates", IEEE PAMI, 23(3):257-267, 2001.

[5] M. Ahad, J. Tan, H. Kim, and S. Ishikawa, "Temporal motion recognition and segmentation approach", Int. J. of Imaging Systems and Technology, 19:91-99, 2009.

[6] H. Bay, A. Ess, T. Tuytelaars, L. Gool, "Speeded-up robust features (SURF)", Computer Vision and Image Understanding, 110(3):346-359, 2008.

[7] M. Fischler and R. Bolles, "Random Sample Consensus: a paradigm for model fitting with applications to image analysis and automated cartography", Communications of ACM, 24(6):381-395, 1981.

[8] M. Ahad, J. Tan, H. Kim, and S. Ishikawa, "Lower-dimensional feature sets for template-based motion recognition approaches", Journal of Computer Science, 6(8):920-927, 2010.