Histogram of multi-directional Gabor filter bank for motion trajectory feature extraction

NGOC NAM BUI, TAN DAT TRINH, MIN KYUNG PARK, JIN YOUNG KIM
Dept. of Electronics and Computer Engineering
Chonnam National University
500-757 Yongbong-dong, Buk-gu, Gwangju
SOUTH KOREA
buingocnam87@gmail.com, trinhtandat1230@yahoo.com, min-0-42@nate.com,
beyondi@chonnam.ac.kr

Abstract: - In this paper, we decompose motion flow using multidirectional Gabor filter bank to create robust descriptors for action recognition problem. The dense trajectory is utilized to allocate the Region of Interest (ROI) for extracting Gabor descriptors. In addition, the feature distribution is adopted instead of the conventional sum of energy. Our experiments are conducted in an open data set (UCF11) and our self-constructed one (CNU). The results indicate that the proposed feature outperforms all other ones in both individually and collectively.

Key-Words: - Gabor histogram, GMM supervector, dense trajectory features.

1 Introduction

Action recognition in recent years has been a challenging problem that attracted profound attention from the research community. The first two benchmarks Weizmann and KTH was introduced by Blank, et al. [1] and Laptev and Lindeberg [2] respectively. Blank, et al. [1] tackle the problem by extracting motion energy based-silhouette image or specifically space-time saliency and orientation characteristics. On the other hand, Laptev and Lindeberg [2] extend their work that identifies two-dimensional local invariant interest point into space-time domain and approached the nearest neighbor searching to classify actions. Generally, those two dataset captures only one movement repeated several times in each sample without any interactions. So we consider them as the ideal datasets that usually gain perfect results. In our previous experiment [3], actions in KTH dataset are recognized with more than 96% accuracy. Later on, UCF 11 or extended Youtube dataset was recorded to represent real-world activities. Each clip is sampled under variations of light and camera motions. Moreover, interaction among candidates, for example spiking ball, contributes significantly to the UCF 11's complexity. The conventional classification which directly classifies the extracted features failed to apply in this case. Therefore, Liu, et al. [4] pursue a Bag of Feature Words that aggregates all descriptors from given video sample into one representative vector. The encoded features are then classified using Support Vector Machine (SVM). Among various feature extraction techniques, Dense Trajectory [5] has been integrated into various systems due to their invention of flexible space-time tube structure wrapping around the trajectory. The structure outperforms the general cuboid that usually accepts a single key point as its centroid. After that, five different descriptors Trajectory position, Histogram of Gradient, Histogram of Flow, Horizontal Motion Boundary Histogram and Vertical Motion Boundary Histogram are extracted to be the trajectory descriptors. Wang, et al. [5] then encode their features using BOF and classify them with SVM under Chi-square kernel. Even though Sobel filter shown its best performance in extracting the motion boundary, it basically strengthens the changes in optical flow following only vertical and horizontal direction. In this paper, we argue that using the derivatives following only x and y directions cannot fully reveal the true motion boundary.

Instead, a multi-directional filter bank with image pyramid is an adequate structure to tackle the issue.

BOW estimates feature distribution using non-parametric approach, otherwise, Perronnin, et al. [6] utilizes the Gaussian mixture distributed assumption to represent the feature space as a Fisher vector. Then the outputs are classified using linear SVM. On the other hand, in our experiment [3], the Gaussian Mixture Model (GMM) Supervector enhanced dramatically the accuracy due to the capability of integrating with various kernels such as RBF, linear/nonlinear GMM Kullback Leibler (GMM KL) and linear/nonlinear GUMI.

2 System overall.

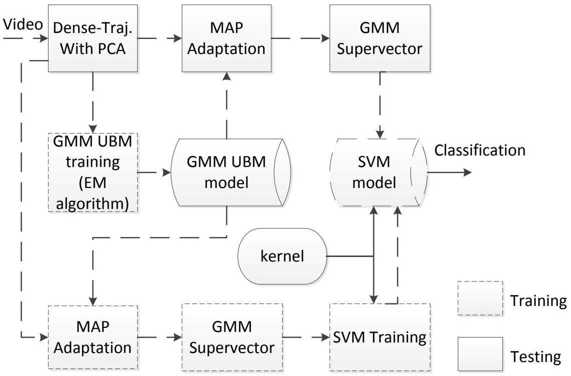


Figure 1: The block diagram of the HAR system-based GMM-UBM supervector.

In our work, we follow the similar scheme of general HAR model that is illustrated in Figure 1. The recognition system using GMM-UBM supervector is described using the block diagram. The training flow is shown in dash line, and the thick one indicates the testing one.

a) Given one sample, the dense trajectory features included traces, HOG, HOF, MBHx and MBHy are extracted. Wang, et al. [5] denotes $n_\sigma x n_\sigma x n_\tau$ is the spatio-temporal size of given grid. In our system, the parameters are chosen as $n_\sigma = 2$ and $n_\tau = 3$.

b) In the next stage, we utilize the GMM-UBM method. All features are pooled to train the UBM model by using an Expectation Maximization (EM) algorithm.

c) After that, the Maximum a Posterior (MAP) Adaptation is utilized to adapt features with UBM model and generate the corresponding GMM model.

d) A mean supervector is obtained from the GMM model by cascading all Gaussian components into a single vector.

e) Finally, the SVM classifier is trained with different kernel functions are utilized to process the classification task.

In our proposed system, features are first projected to the eigenspace using Principal Component Analysis (PCA) to normalize their coordinates which helps to increase the number of mixtures. Because the number of features is enormous, we can only randomly select a part of these features set to train the UBM model. Fortunately, the randomness shows its insignificant impact on the recognition rate.

The SVM classifier is first applied to the binary case. Therefore, in order adjust the SVM to deal with multiple classes' recognition; we adopt an SVM with a one-versus-the-rest technique, which treats the entire class as a single group and the remaining as another one. Various kernels have been developed to be compatible with SVM. However, we only adopt some of the representatives including Linear kernel, nonlinear GMM KL divergence kernel [7], nonlinear GUMI kernel [7] due to their efficiency and widespread usage, as described in [3].

3 Dense Trajectory feature extraction.

Of the various applicable feature extraction techniques, dense trajectories and motion boundary descriptors [5] have shown their outstanding performance and high capability with different data sets. The features points are sampled using grid-based approach for each layer in the scale pyramid to increase the density of the trajectory. The gap between each sample points is set to 5 pixels to maintain the quantity and processing time. Considering the first sample as a root node, then following the optical flow used to describe the pixel movement, the keypoint is traced along the flow direction and to sketch out its trajectory. Finally, a spatial temporal volume, center by each trajectory, is constructed to gather information. Wang, et al. [5] criticize that the essential information about specific action is mostly compacted inside these 3D cubes which later are utilized for extracting descriptors. In summary, the extraction process involves three correlated steps, including sampling interest points and then tracking their trajectory along image sequences. Finally, the local features are obtained by imposing grid method in each 3D trajectory box.

Points are resampled on each spatial scale separately following the scale space theory. The scale pyramid with 8 layers per one octave holds a property of invariance to scale variation but it also leads to a surge in a number of keypoints. In order to compromise between point density and processing time, an adaptive threshold is defined as in Equation (1).

$$T = \rho * \max_{i \in I}\left(\min\left(\lambda_i^1, \lambda_i^2\right)\right), \quad (1)$$

where λ_i^t denotes for the eigenvalue of t direction defined in the Hessian matrix. A point which has large eigenvalues of the auto-correlation matrix tends to be either edge or corner. Thus, the adaptive thresholds, controlled by a single parameter ρ, are used to limit the number of trajectories.

After interest points are all sampled in different scales, the corresponding paths are formed by tracking their location original to the optical flow. Assuming a point P_t located at frame I_t has been detected, its adjacency location in frame I_{t+1} is obtained by Equation (2). Wang, et al. [5] approaches the median filter instead of bilinear interpolation to smooth the optical flow $\omega_t = (u_t, v_t)$ before adding up to P_t. Those trajectories with small length are discarded to reduce the redundant.

$$P_{t+1} = P_t + M * \omega_t \quad (2)$$

Lastly, local descriptors such as HOG, HOF and Motion Boundary Histogram (MBH) are then extracted to give the final presentation.

4 Gabor Filter for Descriptor Representation.

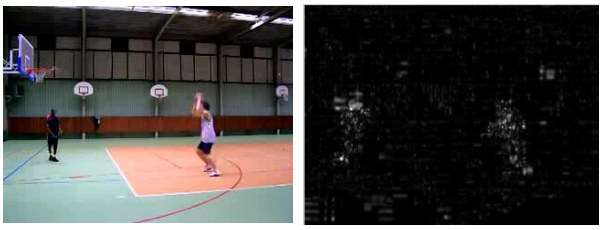
Gabor filter has been widely applied in image processing [6] to imitate human visualization. It's well-known to decompose a single image using a linear combination of different angles and frequencies. Firstly, the pixel candidates are rotated according to the θ angle

$$\bar{x} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} x, \quad (3)$$

where $x = (x, y)$ and $\bar{x} = (\bar{x}, \bar{y})$. Then the Real Gabor function (Equation (4)) is then applied to extract the corresponding feature.

$$\mathcal{G}_{\lambda,\theta,\varphi,\gamma}(x) = e^{-\left(\left(x'^2 + \gamma^2 y'^2\right)/2\sigma^2\right)} \cos\left(2\pi \frac{x'}{\lambda} + \varphi\right) \quad (4)$$

In this formula, λ represents the scale or frequency, φ is the initial phase and γ is the spatial aspect ratio. The imaginary form of this filter can be obtained by simply replacing the cosine with the sine function. In this paper, instead of applying the Gabor kernel directly to the image, we use the filter in the flow field with the purpose of revealing the different direction of movements. Given a particular frame in basketball match video as shown in**Error! Reference source not found.** (a), we demonstrate its average energy of Gabor flow and the corresponding components in**Error! Reference source not found.** (b), and **Error! Reference source not found.** respectively. As can be observed, the intensive motion areas after the filtering process respond significantly higher than their neighbours.



(a) (b)

Figure 2: Illustration of Gabor average energy in optical flow field in basketball video of UCF11 dataset.

(a) Original video

(b) the average energy image of (a).

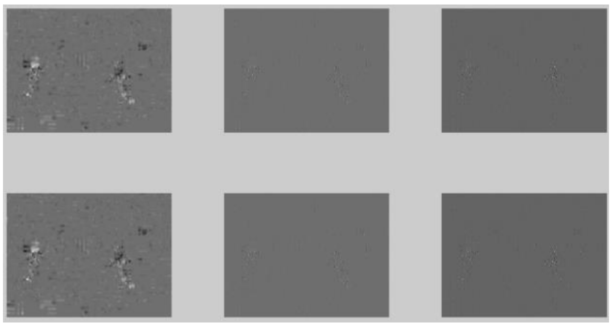


Figure 3: The image energy of six components corresponding to different directions of Gabor features.

Rather than applying the phase or energy of each component to shape the feature form, we argue that the energy distribution of each Gabor feature might contain more information. Each Gabor is represented by a complex number which is effective to build the Histogram of Orientation. Given $I_{\mathcal{G}}^{re}$ and $I_{\mathcal{G}}^{im}$ is the real and imaginary part of Gabor component, the concept described above can be formalized as:

$$h_{b_i}(\theta_{x,y}) = h_{b_i}(\theta_{x,y}) + m \quad (5)$$
$$\text{if} \quad \theta_{x,y} \in b_i,$$

where $\theta = \arctan\frac{I_{\mathcal{G}}^{im}}{I_{\mathcal{G}}^{re}}$ and $m = \sqrt{\left(I_{\mathcal{G}}^{im}\right)^2 + \left(I_{\mathcal{G}}^{re}\right)^2}$ is denoted as the angle and magnitude of each Gabor Image pixel.

5 Gaussian Mixture Model Supervector.

The GMM model is represented by a set of parameters $\lambda = \{\lambda_m\} = \{\omega_m, \mu_m, \Sigma_m\}, i = 1, \dots, M$ where ω_m, μ_m and Σ_m are the weight, mean and covariance of an $m-th$ component respectively. Given a sample x, its probability is calculated by the weighted sum of M Gaussian components.

$$p(x|\lambda) = \sum_{i=1}^{M} \omega_m g(x|\mu_m, \Sigma_m), \quad (6)$$

where as $g(x|\mu_m, \Sigma_m)$ is the Gaussian function of m-th mixture defined in Equation (7).

$$g(x|\mu_m, \Sigma_m) = \frac{1}{\sqrt{2\pi}|\Sigma_k|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x - \mu_k)'\Sigma_k^{-1}(x - \mu_k)\right\} \quad (7)$$

The mixture of mean values and covariance is then cascaded to give the final representation named as GMM supervector. The feature vectors are classified using SVM with Linear, nonlinear GUMI and nonlinear GMMKL kernel.

$$k_{\text{nonlinearKL}}(g^a||g^b) = \sum_{i=1}^{M} e^{-\frac{\left\|P_i^a - P_i^b\right\|}{2\sigma^2}}, \quad (8)$$

where $P_i^a = \sqrt{\omega_i}(\Sigma_i^u)^{-1/2}\mu_i^a$ and $P_i^b = \sqrt{\omega_i}(\Sigma_i^u)^{-1/2}\mu_i^b$.

$$k_{\text{nonlinearGUMI}}(g^a||g^b) = \sum_{i=1}^{M} e^{-\frac{\left\|S_i^a - S_i^b\right\|}{2\sigma^2}}, \quad (9)$$

where

$$S_i^a = \left(\frac{\Sigma_m^a + \Sigma_m^u}{2}\right)^{-1/2} (\mu_m^a - \mu_m^u) \text{ and}$$

$$S_i^b = \left(\frac{\Sigma_m^b + \Sigma_m^u}{2}\right)^{-1/2} (\mu_m^b - \mu_m^u).$$

6 Dataset and Experimental Results.

We evaluate our proposed features using an open dataset UCF11 and our self-recorded one CNU. The UCF11 so far has been recommended to one of the most complicated benchmarks that is variance regarding illumination changes, camera angle variation, object scales, viewpoint and object interactions. However, it mostly focuses on sport actions such as tennis swing, jumping, cycling or walking. Then, to serve for our particular purpose which aims to prevent the violent behaviour in university, we recorded our CNU dataset that focusing on two actions set: fighting and the others. The data collection has been conducted at our school and surrounding areas while ensuring the realistic of human under various types of scenarios. For each condition, we collected the information from two groups, males, and females, performing various actions. Moreover, the school environment requires our method to overcome the similarity causing by student uniform. In several groups such as drinking fountain, the scale of targets is diversified dramatically. In details, our dataset is acquired in different places as shown in Figure 2.30. Then each action clip is manually cropped to provide the total of 455 fighting activities and 399 non-fighting ones. We take 50% of data used for training, and the remaining is for testing. In general, our setup consists of 236 fighting and 207 non-fighting clips for training versus 219 clips and 192 clips for a testing fight and non-fight action. On the other hands, the UCF11 experiments are proceeded using leave-one-out cross-validation with 25 folds as mentioned in [4].

Table 1 shows the performance of different descriptors with three kernels. As can be observed, Gaborflow feature outperforms the other in all of three cases. Specifically, the nonlinear GUMI acquires the best accuracy rate with 75.87% which is 2.78% better than the MBHy within the same kernel.

Table 1: Performance of Different Features on CNU Dataset.

Kernel	8 mixtures					
	Traj. Pos.	*HOG*	*HOF*	*MBHx*	*MBHy*	*Gaborflow*
Linear	57.42	64.62	68.45	65.31	65.20	**69.37**
Nonlinear GUMI	66.94	70.42	73.43	73.09	72.27	**75.87**
Nonlinear GMMKL	67.63	70.42	74.13	73.78	72.16	75.29

We also integrate the new features with other ones that slightly boost the performance to 77.96%.

Table 2: Experiments of various channel combination in CNU Dataset.

Kernel	8 mixtures	
	(Traj.Pos.+HOG+HOF +MBHx+MBHy)	*(Traj.+HOG+HOF+MBHx+MBHy)+Gaborflow*
Linear	71.69	73.09
Nonlinear GUMI	77.49	77.96
Nonlinear GMMKL	77.61	77.61

In the following part, the open dataset UCF11 is verified using similar scheme with only four mixtures of Gaussian distribution. The results are shown in Table 3 and Table 4. The Gabor feature still reaches the highest rate except the nonlinear GUMI. In the case of the combinations, the maximum rate (80.27%) is obtained with the nonlinear GMMKL.

Table 3: Performance of Different Features on UCF11 Dataset.

Kernel	8 mixtures					
	Traj. Pos.	*HOG*	*HOF*	*MBHx*	*MBHy*	*Gaborflow*
Linear	21.98	60.28	57.52	54.30	51.11	60.60
Nonlinear GUMI	33.86	62.06	59.65	60.39	57.79	60.92
Nonlinear GMMKL	39.46	71.18	69.57	68.99	66.39	72.16

Table 4: Experiments of various channel combination in UCF11 Dataset.

Kernel	8 mixtures	
	(Traj.Pos.+HOG+HOF +MBHx+MBHy)	*(Traj.+HOG+HOF+MBHx+MBHy)+Gaborflow*
Linear	71.69	73.09
Nonlinear GUMI	77.49	77.96
Nonlinear GMMKL	77.61	77.61

Table 5: Comparison with Latest Approaches.

Perez, et al. [8]	Hasan and Roy-Chowdhury [9]	Mota, et al. [10]	Our method
68.9%	69.0%	72.7%	80.27%

7 Conclusion

In this paper, we have demonstrated an application of Gabor filter bank on optical flow which helps to generate a distinct feature. The experimental results indicate that our features achieve the best score among all representatives. Moreover, the combination of all elements with nonlinear GMMKL has exceeded other recent methods. However, the extracting procedure does consume much time that is required for further enhancement for online classification purpose.

Acknowledgement

This research was supported by the Ministry of Education, Science Technology (MEST) and National Research Foundation of Korea(NRF) through the Human Resource Training Project for Regional Innovation.

References:

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as Space-Time Shapes," *Proceeding of Tenth IEEE International Conference on Computer Vision*, Beijing, vol. 2, no., pp. 1395 - 1402, Oct. 2005 2005.

[2] I. Laptev and T. Lindeberg, "Local Descriptors for Spatio-temporal Recognition," *Proceeding of First International Workshop, SCVMA*, vol. 3667, no., pp. 91-103, 2004.

[3] N. N. Bui and J. Y. Kim, "Human Action Recognition based on GMM-UBM Supervector using SVM with non-linear GMM KL and GUMI," *Proceeding of Int. Conf. Digital Image Processing*, vol. 9631, no., pp. 96311G-96311G-7, April 2015.

[4] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1996 - 2003, 2009.

[5] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense Trajectories and Motion Boundary Descriptors for Action Recognition," *International*

Journal of Computer Vision, vol. 103, no. 1, pp. 60-79, May 2013.

[6] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher Kernel for Large-Scale Image Classification," *Proceeding of Computer Vision – ECCV,* vol. 6314, no., pp. 143-156, January 2010.

[7] N. N. Bui, J. Y. Kim, and T. D. Trinh, "A non-linear GMM KL and GUMI kernel for SVM using GMM-UBM supervector in home acoustic event classification," *The Institute of Electronics, Information and Communication Engineers.,* vol. E97-A, no. 8, pp. 1791-1794, August, 2014.

[8] E. A. Perez, V. F. Mota, L. M. Maciel, D. Sad, and M. B. Vieira, "Combining gradient histograms using orientation tensors for human action recognition," *Proceeding of 21st International Conference on Pattern Recognition (ICPR),* pp. 3460-3463, 11-15 Nov. 2012 2012.

[9] M. Hasan and A. K. Roy-Chowdhury, "Incremental Activity Modeling and Recognition in Streaming Videos," *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 796-803, 23-28 June 2014 2014.

[10] V. F. Mota, E. A. Perez, L. M. Maciel, M. B. Vieira, and P. H. Gosselin, "A tensor motion descriptor based on histograms of gradients and optical flow," *Pattern Recognition Letters,* vol. 39, no., pp. 85-91, 4/1/ 2014.