Exploring Pooling Strategies based on Idiosyncrasies of Spatio-Temporal Interest Points

Yuancheng Ye, Xiaodong Yang, and Yingli Tian
The Graduate Center and The City College
City University of New York
yye@gradcenter.cuny.edu, {xyang02, ytian}@ccny.cuny.edu

ABSTRACT

Recent studies have demonstrated that the implementation of local space-time interest points has good competence and robustness in the area of human action recognition, which has become one of the challenging problems in multimedia analysis. While most research focuses on the techniques of detecting feature points or capturing spatial and temporal information around those points, there has been very limited research on delving into the pooling strategies which are also important components of action recognition algorithms. In this paper, we propose a novel pooling framework by categorizing the interest points with respect to their idiosyncrasies. Specifically, we discuss three pooling strategies based on the optical flow orientation, foreground weight and spatio-temporal locations respectively and further investigate the fusion of different pooling strategies. For the encoding process, instead of the popular bag-of-visual words (BoV) method, we adopt the improved Fisher Vector (FV) approach. Our proposed methods are evaluated on a benchmark dataset with controlled settings (KTH), and two more challenging datasets with realistic background (HMDB51 and UCF101). The experimental results demonstrate that pooling strategies based on the appropriate idiosyncrasies of individual interest points can improve the performance of action classification.

Keywords

Action recognition, local interest points, pooling strategy, spatio-temporal pyramid, optical flow, foreground weight

1. INTRODUCTION

Recent years, action recognition has drawn more and more attentions in the area of multimedia analysis. Among many approaches developed to solve this challenging problem, local space-time features have demonstrated good robustness and competence since they can describe relatively independent representation of actions with respect to their spatio-temporal shifts. These features are usually detected directly from videos and therefore without pre-processing procedures such as motion segmentation and tracking [30].

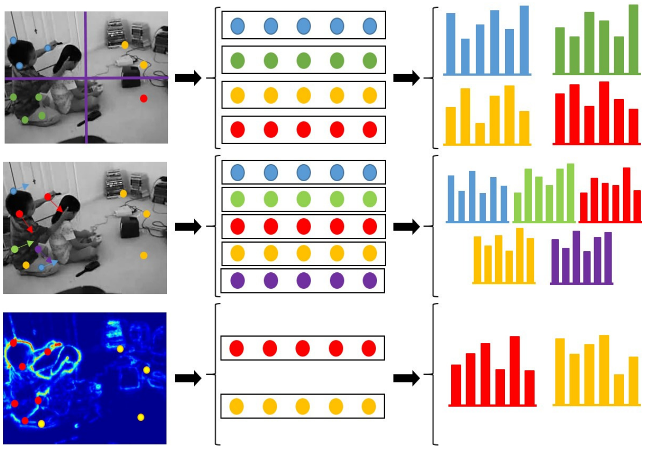

Figure 1: An illustration of three pooling strategies we evaluate in this paper: spatio-temporal pyramid pooling strategy (top row), the optical flow orientation-based pooling strategy (middle row) and foreground weight-based pooling strategy (bottom row). In the first column, the circles in the figure represent different detected interest points, and different colors represent the distinct properties associated with them. Second column shows the procedure of the interest points categorized according to their idiosyncrasies. Third column illustrates the vectors of different channels after sum pooling and normalization procedures respectively.

1.1 Related Work

Various local feature detectors have been proposed in the past years. These detectors are usually obtained by maximizing a certain saliency function with respect to given spatio-temporal scales.

By extending the Harris corner detector to the 3D domain, Laptev and Lindeberg first proposed the STIPs for action recognition [14] by extracting sparse interest points in the space-time domain. Cuboid detector [6] and 3D-Hessain detector [31] were proposed by Dollár *et al.* and Willems *et al.* respectively. Chakraborty *et al.* proposed selective spatio-temporal interest points [5]. By employing state-of-art optical flow fields, Wang *et al.* [28, 29] introduced dense

ICMR '15, June 23 - 26, 2015, Shanghai, China
Copyright 2015 ACM. ISBN 978-1-4503-3274-3/15/06 ...$15.00
DOI: http://dx.doi.org/10.1145/2671188.2749339.

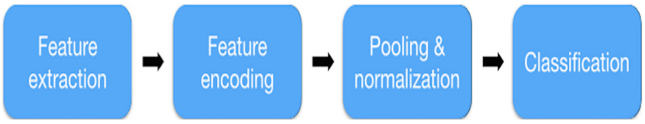

Figure 2: A traditional pipeline of action recognition algorithms instead of what this paper presents.

trajectories.

To capture the unique appearance and motion information around interest points, a number of descriptors have been proposed [21]. Schüldt *et al.* [23] tried to tackle action recognition problems by applying high-order derivatives to compute local N-jets descriptors. Laptev [15] introduced HOG/HOF based on the distribution of image gradients and optical flow around the neighborhood of interest points. Dollár and his colleagues proposed cuboid descriptor along with the cuboid feature detector [6]. 3D-SIFT [24] and HOG3D [12] descriptors both were introduced by extending successful 2D image descriptors into 3D domain. In particular, the paper [30] conducted exhaustive experiments to evaluate the performance of different feature detectors and feature descriptors.

In the literatures of human action recognition, bag-of-visual words (BOV) method is a common encoding strategy to compute the representation of action videos. It entails codebook vocabularies usually computed by the K-means algorithm. The number of clusters is set empirically to obtain the best results. Additionally, to achieve a good classification performance, a non-linear support vector machine (SVM) is usually necessary. However as underlined in the work of Boiman *et al.*, the descriptor quantization is a lossy process which is one of the main shortcomings of the BOV representation [2]. Recent research has demonstrated that the improved Fisher Vector (FV) encoding approach consistently outperforms the traditional BOV method [22] by implementing the ℓ_2 normalization and power normalization [11]. The FV representation works well with linear SVM which makes the action recognition tasks possible for large-scale datasets while achieving comparable or even better results than the BOV encoding approach with non-linear SVM.

Pooling process often follows the encoding procedure in action recognition algorithms [17]. Two aggregating strategies are commonly used: sum pooling and max pooling. The paper [3] conducted an extensive theoretical analysis of feature pooling strategies, and as it pointed out that sparse features prefer max pooling approach [20]. These two pooling strategies are performed on a global scale, therefore may neglect the discriminative properties of intra-descriptors in the same video. In our paper, instead of pooling all descriptors without prejudice, we prove that pooling strategies based on the idiosyncrasies of local interest points can improve the discriminative ability of video representations.

1.2 Our Proposed Framework

As discussed above, a classical action recognition algorithm usually comprises of four main components: feature extraction, feature encoding, pooling and normalization, and classification as shown in Figure 2. Although many studies have been done on feature extraction and feature encoding, there is limited research delving into the pooling strategies based on the idiosyncrasies of individual interest points.

Most action recognition algorithms employ sum pooling or max pooling strategy depending on the encoding process implemented. Specifically, sparse coding employs max pooling while traditional BoV encoding algorithm uses sum pooling. These two standard pooling strategies are based on the elements of descriptors and they are more likely regarded as the integrating procedures. However, as demonstrated by our experiments, the discernability provided by appropriate properties-based pooling strategies can have positive influence on the classification performance. As one of the most commonly implemented pooling strategies based on the properties of interest points, spatio-temporal pyramid pooling method [16], which can be shown to be a special pooling case in our framework, categorizes detected feature points by their space-time positions in video clips. Many previous studies have demonstrated that this approach conspicuously improves recognition results. Most literature regards the spatio-temporal pyramid pooling strategy merely as a special implementation of pyramid match kernel [7]. However we argue that it can also be regarded as a specific pooling strategy based on the positions of interest points.

Inspired by the spatio-temporal pyramid pooling strategy, which capitalizes the location information of individual interest points, we propose two other idiosyncrasies-based pooling strategies to improve the discernability of video representations: the optical flow orientation-based and foreground weight-based pooling strategies. The purpose of the first pooling strategy is to utilize the optical flow orientation information, while the second is to decide whether the point belonging to the foreground scenes. Figure 1 illustrates the three pooling strategies we discussed in this paper. As with the location information, we argue that optical flow orientation and foreground weight information of interest points are also informative properties that could be applied to pooling procedures.

In addition to the specific pooling strategies we evaluate in this paper, we generalize a new pooling framework based on the idiosyncrasies of individual interest points. In our new framework, interest points are pooled according to their distinct properties. As demonstrated by our experiments, more discriminative power can be rendered to the representation by implementing our pooling framework.

1.3 Contributions and Paper Organization

In this paper, we propose a novel pooling framework by exploring the idiosyncrasies of individual interest points. Specifically, we study two novel pooling strategies based on the interest points properties of optical flow orientation and foreground weight respectively and further integrate these two new pooling methods with the popular spatio-temporal pyramid strategy to achieve better recognition results on different datasets of human action recognition. As presented in Figure 3, we first conduct three pooling strategies respectively, and then to integrate the performance of these pooling strategies, we employ a decision-level fusion to further improve the results. Our experiments demonstrate that pooling strategies based on categorizing interest points with respect to their appropriate and reasonable properties can improve the action recognition performance.

The rest of the paper is organized as follows. Section 2 discusses the general idiosyncrasies-based pooling framework and two novel pooling strategies: the optical flow orientation-

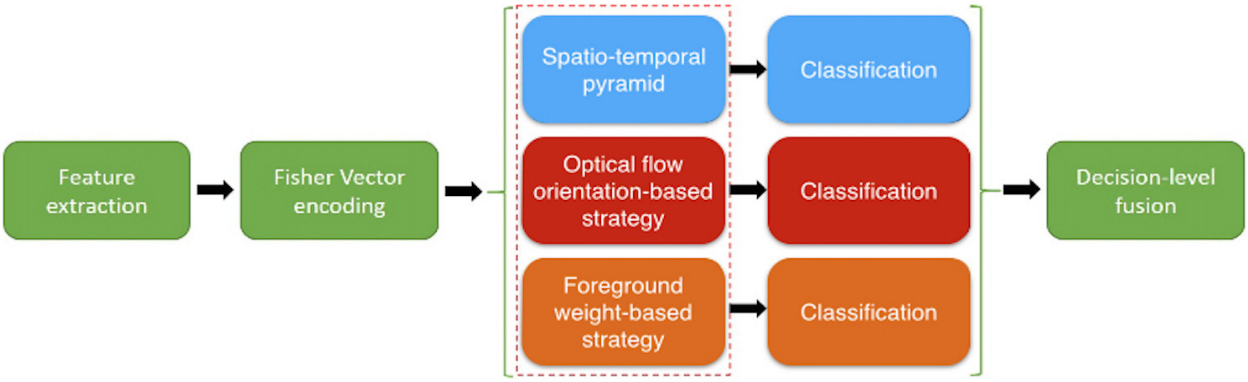

Figure 3: The pipeline of our proposed methods. The main contribution of this paper is denoted by the dashed box which represents our proposed idiosyncrasies of interest points-based pooling framework. Spatio-temporal pyramid, optical flow orientation-based strategy and foreground weight-based strategy are all the specific examples of our new pooling framework.

based and foreground weight-based methods. Spatio-temporal pyramid pooling strategy is also briefly discussed from a different point of view. Section 3 presents the experimental setup details and results on different datasets. We conclude the paper in Section 4.

2. IDIOSYNCRASIES OF INTEREST POINTS-BASED POOLING STRATEGIES

Two popular encoding frameworks are commonly employed in action recognition algorithms: BOV words and Fisher Vector (FV) [32]. Recent research has demonstrated that the Fisher Vector encoding procedure consistently achieves more accurate results with appropriate setups. In order to clearly present our proposed pooling framework, we firstly simply introduce the FV encoding process.

Let $\chi = \{x_1, x_2, \ldots, x_T\}$ be the descriptors in an action video, assuming our GMMs have the form:

$$u_\lambda(x_t) = \sum_{j=1}^{K} \omega_j u_j(x_t), \qquad t = 1, 2, \ldots, T, \tag{1}$$

in which $\lambda = \{\omega_i, \mu_i, \Sigma_i, i = 1 \ldots K\}$, and ω_i, μ_i and Σ_i are the mixture weight, mean vector and covariance matrix of Gaussian u_i respectively. Here K is the number of the GMMs components. We also assume that the covariance matrices are diagonal and then denote the variance vector by σ_i^2. $\gamma_t(i)$ is the soft assignment of descriptor x_t to Gaussian component i:

$$\gamma_t(i) = \frac{\omega_i u_i(x_t)}{\sum_{j=1}^{K} \omega_j u_j(x_t)}. \tag{2}$$

Encoded vector of x_t is then given by the concatenation of vectors $\Theta_{\mu,i}^t$ and $\Theta_{\sigma,i}^t$, which are defined as follows:

$$\Theta_{\mu,i}^t = \frac{1}{\sqrt{\omega_i}} \gamma_t(i) \left(\frac{x_t - \mu_i}{\sigma_i} \right), \tag{3}$$

$$\Theta_{\sigma,i}^t = \frac{1}{\sqrt{2\omega_i}} \gamma_t(i) \left(\frac{(x_t - \mu_i)^2}{\sigma_i^2} - 1 \right). \tag{4}$$

The concatenation process can then be expressed as:

$$\rho(x_t) = [\Theta_{\mu,1}^t, \Theta_{\sigma,1}^t, \ldots, \Theta_{\mu,K}^t, \Theta_{\sigma,K}^t]. \tag{5}$$

Suppose the dimension of the descriptor x_i is D, then the encoded descriptor $\rho(x_t)$ has the size $2DK$.

After the encoding procedure, an appropriate pooling strategy should be applied. The standard pooling approach is the sum pooling procedure. Then the power normalization and ℓ_2-normalization are performed to obtain the final representation. Since sum pooling strategy is based on the elements of vectors respectively on a global scale, it does not discriminate descriptors in the same video. Therefore it may neglect some discernabilities intrinsic among descriptors. To underline the different properties of interest points, spatio-pyramid pooling strategy is usually applied before standard sum pooling process.

Grauman *et al.* first introduced pyramid kernel to classify different features [7]. Inspired by this work, Lazebnik *et al.* [16] proposed a simple but effective spatial pyramid technique of pooling feature points into different bins based on their positions in an image. By extending this pooling strategy into 3D domain, the spatio-temporal pyramid pooling strategy was employed in the paper [27]. This approach consistently improves recognition performance on different datasets.

Inspired by the success of spatio-pyramid pooling strategy, we propose a general pooling framework in this paper which projects interest points based on their distinct properties. Let $\eta = \{y_1, y_2, \ldots, y_T\}$ be the set of locations of detected interest points. Assuming the projection procedure is denoted as φ, the mathematical expression of our idiosyncrasies-based pooling process is defined as:

$$C_n = \varphi(\{y_t\}), \tag{6}$$

where C_n is an element of the pooling set $\{C_1, C_2, \ldots, C_N\}$ whose components are the feature indices. The aggregated vector h_n belonging to the pooling component C_n then can be expressed as:

$$h_n = \frac{1}{|C_n|} \sum_{x_t \in C_n} \rho(x_t). \tag{7}$$

Finally, the normalized aggregated vectors are concatenated to form the video representation:

$$z = [h'_1, h'_2, \ldots, h'_N]^T. \tag{8}$$

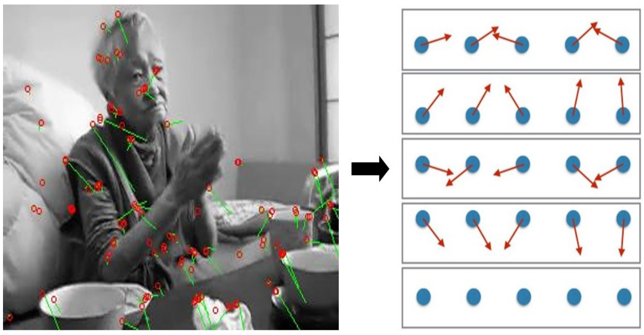

Figure 4: An intuitive illustration of optical flow orientation-based pooling strategy. According to their optical flow orientation information, interest points are projected into five categories: $-90^o \sim -45^o, -45^o \sim 0^o, 0^o \sim 45^o, 45^o \sim 90^o$ and *static*.

where $h'_n, n = 1, 2, \ldots, N$, denotes the normalized and transposed form of the aggregated vector h_n.

One main contribution of this paper is to evaluate the effectiveness of the discernability of video representation by carefully applying reasonable pooling procedure φ. In our proposed framework, interest points are categorized according to their properties, and then encoded descriptors are computed respectively. After that standard pooling and normalization procedures are operated on each category. Finally a representation is formed by concatenating vectors of different categories.

Like the spatio-temporal pyramid pooling strategy, the pooling procedure φ groups interest points into different categories and therefore increases the discriminative power of final representation. As implementations of our proposed pooling framework based on the idiosyncrasies of interest points, two novel pooling strategies are presented in our paper: optical flow orientation-based and foreground weight-based pooling strategies.

2.1 Optical Flow Orientation-based Pooling Strategy

Optical flow is commonly used for tracking, motion estimation, etc. The paper [1] discussed several most commonly used optical flow algorithms. Since optical flow encodes the motion information of points in a frame relative to the next frames, it is an intuitive source to describe action motion. We incorporate the optical flow orientation information in the pooling process φ to project interest points into different categories.

As illustrated in Figure 4, the detected interest points are associated with different optical flow angles. These optical flow angles indicate the orientations where those interest points will be in the next frame. Interest points with different optical flow angles contain different information. By pooling interest points according to optical flow orientations, more discriminative information is represented for video actions.

At first, by discarding the interest points without optical flow value, we compute the optical flow angles by the *arctan* function of the projections of amplitudes on X-coordinate and Y-coordinate, and divide the angles into four categories: $-90^o \sim -45^o, -45^o \sim 0^o, 0^o \sim 45^o, 45^o \sim 90^o$. The categorization process is the implementation of the pooling procedure φ in Eq.(6). However, results on action recognition are even worse than the baseline. The main reason is that, as shown in Figure 4, there are many interest points associated with relatively small or even zero amplitudes of optical flow. A straightforward intuition is that these interest points are on the background settings and should be eliminated for further action recognition. However the experiment results do not support this conjecture.

Actually there are many existing literatures discussing the discernability of background scenes. Furthermore, those static interest points may also be generated by temporary action pause which still provide rich information about action movement. Neglecting these points will inevitably deteriorate the performance. Therefore, we add a *static* category to accommodate for this situation, which helps to improve the performance. This phenomenon of many detected interest points falling in the static background settings inspires us to separate the feature points into foreground and background bins, which will be discussed in next Section.

2.2 Foreground Weight-based Pooling Strategy

Recently, an interesting experiment discussing the background discriminativity demonstrated that comparable recognition results can be achieved by dealing with the features points extracted only on the background regions [26]. Inspired by this work, we employ the value of foreground weight to measure the confidence of how likely a point belongs to the foreground scenes based on the information of optical flow gradients, color gradients, and visual saliency.

The optical flow gradients-based foreground confidence, f_m, can be calculated as follows:

$$f_m(x, y) = \sqrt{\mu_x^2 + \mu_y^2 + \upsilon_x^2 + \upsilon_y^2} * g, \tag{9}$$

where μ_x and υ_x are the gradients of the projection of optical flow magnitude on the X-coordinate, and correspondingly μ_y and υ_y are the gradients on the Y-coordinate. g represents the 2D Gausian filter with a fixated variance and the symbol $*$ denotes the convolution process. A visualization of the confidence map computed by optical flow gradients is shown in Figure 5(a).

The Frobenius form of the color gradients, f_c, is:

$$f_c(x, y) = \sqrt{L_x^2 + L_y^2 + a_x^2 + a_y^2 + b_x^2 + b_y^2} * g, \tag{10}$$

where (L_x, a_x, b_x) and (L_y, a_y, b_y) are the horizontal and vertical color gradients of the pixel at the position (x, y). Figure 5(b) presents the heatmap of the foreground confidence map calculated by color gradients.

For the visual saliency, we implement the graph-based visual saliency (GBVS) algorithm, which was originally proposed by Harel [8].

Assuming that the dissimilarity of two points $M(i, j)$ and $M(p, q)$ is defined as:

$$d((i, j) \parallel (p, q)) = \left| log \frac{M(i, j)}{M(p, q)} \right|. \tag{11}$$

Considering all pixels in a frame as connecting nodes in a graph, the GBVS algorithm attempts to assign a weight to every directed edge from the node (i, j) to the node (p, q):

$$w((i, j) \parallel (p, q)) = d((i, j) \parallel (p, q)) \cdot F(i - p, j - q). \tag{12}$$

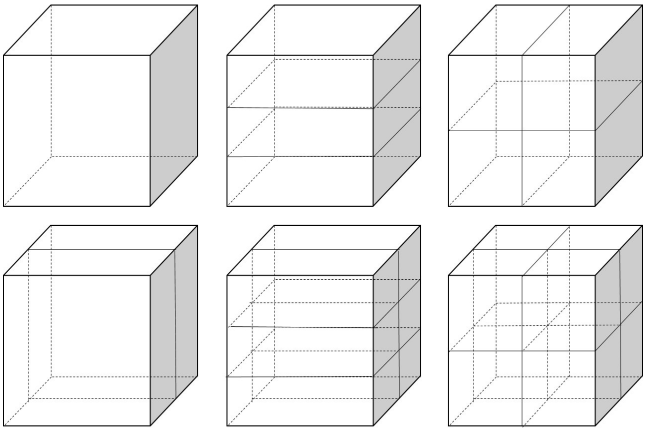

Figure 6: An illustration of spatio-temporal grids distributions in our experiments. Top row: $1 \times 1 \times 1$, $1 \times 3 \times 1$, $2 \times 2 \times 1$. **Bottom row:** $1 \times 1 \times 2$, $1 \times 3 \times 2$ **and** $2 \times 2 \times 2$. **Total of 24 grids are used.**

The function $F(a, b)$ is defined as below:

$$F(a, b) = exp\left(-\frac{a^2 + b^2}{2\sigma^2}\right), \tag{13}$$

where σ is a user defined parameter, usually in the domain of one tenth to one fifth of the video width.

After assigning every edge a saliency weight, the outbound edges are then normalized and the graph is treated as a Markov chain. The equilibrium distribution of this chain can be computed as the saliency measure, $f_s(x, y)$, of each pixel. A visualization of visual saliency heatmap is demonstrated by Figure 5(c).

By incorporating all of three confidence weights, the final foreground confidence map can be obtained:

$$f_{conf} = log(f_m(f_c + f_s) + 1). \tag{14}$$

Figure 5(d) shows the heatmap of final foreground confidence map.

The foreground weight-based pooling strategy is an implementation of our pooling framework φ. It takes into account the discernability of background and foreground by deciding whether an interest point belonging to the foreground scenes to improve the discriminative power of video representations.

2.3 Spatio-Temporal Pyramid-based Pooling Strategy

As a successful pooling strategy, spatio-temporal pyramid has been used in many existing action recognition algorithms. It groups interest points according to their positions in action videos belonging into corresponding geometric grids and calculate their descriptors respectively. Aggregated vectors are obtained after conducting pooling and normalization process on each category. The concatenation of these vectors is presented as the final representation. Although spatio-pyramid approach is evolved from pyramid classification kernel, it can be regarded as a specific example of the idiosyncrasies-based pooling strategies which categorizes interest points with respect to their geometric information. Therefore by applying other reasonable properties of interest points, recognition performance can also be improved due to more discriminative power of representations.

Methods	baseline	4-bins	5-bins
Lucas–Kanade	92.7	91.3	92.2
Horn–Schunck	92.7	92.4	93.2

Table 1: Average accuracies (%) of optical flow orientation-based pooling strategies on the KTH dataset. Baseline column represents the results with standard Fisher Vector method.The 5-bins column presents the results of our final approach which projects STIPs into four angle bins and one *static* bin, while the column of 4-bins lists the outcomes only with four angle bins.

In our paper, we also evaluate how the other two properties, i.e., optical flow orientation and foreground weight, can be applied to our proposed pooling framework to improve the action recognition results. In addition to be an integrating process, the pooling procedure φ, can also improve the discernability of representations by applying appropriate properties.

3. EXPERIMENTS

3.1 Experimental Setup

Our experiments focus on the pooling strategies based on the distinct properties of individual detected interest points. In our implementation, we employ the STIPs developed by Laptev *et al.* as the interest points and the popular HOG/HOF as the descriptors.

HOG/HOF descriptor divides the neighborhood of each interest point into 18 grids and for each grid 4-bin histograms of gradient orientations (HOG) and 5-bin histograms of optical flow (HOF) are computed. Therefore for each interest point, the dimension of the descriptor is 162. After that, a principal component analysis (PCA) procedure is performed for the purpose of decorrelation. The dimension of PCA matrix is set to reserve 98% energy of the original descriptor, which reduces primitive vector size of HOG/HOF from 162 to 126 for all the datasets. Our Gaussian Mixture Models (GMMs) are trained using the VLFeat library and the number of components K is set to 256. Therefore for each interest point, a Fisher Vector is formed with dimension $2 \times 256 \times 126 = 64512$.

3.1.1 Datasets

Experiments have been conducted on the following datasets:

KTH dataset [23] consists of six action classes: walking, jogging, running, boxing, waving and clapping. There are total of 2391 videos under controlled background settings. Following the original experimental setup of the authors, video samples are divided into training set and test set according to different subjects.

HMDB51 dataset [13] is collected from various sources. It consists of 51 action categories and 6766 video sequences. We follow the original protocol using three train-test splits. For each action class, there are 70 video samples for training and 30 video samples for testing. This dataset provides both original and stabilized videos, our results are based on the original videos.

UCF101 dataset [25] is the newest and largest dataset

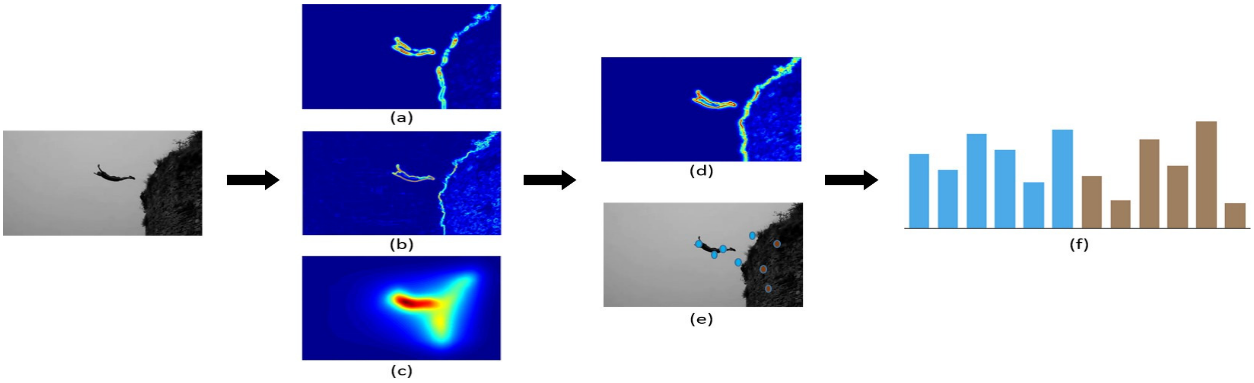

Figure 5: The illustrative pipeline of foreground weight-based pooling strategy. (a) The confidence heatmap computed by optical flow gradients, while (b) and (c) are computed from color gradients and visual saliency respectively. (d) The final foreground confidence map. (e) An illustration of detected interest points in the foreground (blue points) and background (brown points) scenes. (f) The concatenation of encoded vectors of interest points pooling into the foreground and background bins respectively.

which is the extension of the UCF50 dataset. It contains 101 action classes that can be divided into five types: human-object interaction, body-motion, human-human interaction, playing musical instruments, and sports. There are total of 13320 video clips, with fixed frame rate of 25 FPS and resolution of 320 × 240. Our algorithm is evaluated based on the three train/test splits provided by the authors.

3.1.2 *Optical Flow Orientation-based Pooling Strategy*

For the optical flow orientation-based method, we compare two optical flow algorithms on the KTH dataset: 1) Horn–Schunck [9] and 2) Lucas–Kanade [19]. As observed in Table 1, the outcomes with Horn–Schunck algorithm outperform that with Lucas–Kanade method. The reason may be that Lucas–Kanade method is a local approach, the optical flow computed for each interest point can differ very much from others. In contrast, the Horn–Schunck optical flow algorithm is a global approach and the optical flow information of each interest point is relatively continuous with each other. Because STIPs extracted by feature detectors are sparsely scattered in the whole video, the global Horn–Schunck optical flow algorithm is more suitable for our algorithm.

At first we divide the optical flow angles into four bins and categorize detected STIPs accordingly. However the performance decreases. As discussed in the Section 2.1, the static STIPs also provide discernable information. By only using four bins, many STIPs with relative small or even zero optical flow magnitudes are very likely ignored. Discarding this information inevitably results in inferior outcomes.

To accommodate for those static STIPs, we categorize the interest points into five bins by adding a *static* category: $-90^o \sim -45^o, -45^o \sim 0^o, 0^o \sim 45^o, 45^o \sim 90^o$ and *static*. Fisher Vectors are computed for each interest points and then standard sum pooling is performed on each bin. After that power normalization and ℓ_2 normalization are conducted to these five vectors respectively. Finally they are concatenated into one representation with a dimension of 322560. Linear SVMs are used as the classifiers. Our experiments demonstrate that by categorizing the STIPs according to the angles of optical flow, the recognition performance can be improved, which confirms our primitive conjecture about pooling strategies.

3.1.3 *Foreground Weight-based Pooling Strategy*

In addition to the optical flow orientation-based pooling strategy, we also propose a foreground weight-based pooling strategy inspired by [26]. According to Cao *et al.* [4], there are only 18% of all the STIPs detected by Laptev's detector corresponding to the actions performed in the MSR I dataset, while the rest of the STIPs belong to the background. However, based on the discussion in [26], the information of STIPs associated with the background can also benefit action classification. For example the action of riding a bike, the background of a road is most likely associated with the action. Indeed, the background discriminativity can achieve comparable performance as well, which inspires us to categorize the STIPs into foreground and background bins respectively.

We compute the foreground weight by taking into account of motion gradients, color gradients and visual saliency. To cancel out the effect of camera motion, instead of directly using optical flow magnitudes, we compute the derivatives of the horizontal and vertical amplitudes of optical flow respectively, and then the Frobenius form of these gradients. Color information is also an important cue for the foreground. In this paper, we employ the LAB color space to compute the color gradients of each pixel. Then visual saliency is conducted to compute the areas capturing most human attentions. All these three kinds of information are integrated to obtain the final foreground confidence map f_{conf}.

Unlike the approach employed by [26], which applied max-normalization directly to the confidence map f_{conf}, we perform power normalization prior to the max-normalization. This technique is also employed in [22] to reduce the peak phenomenon and improve the performance of Fisher Kernel. Figure 5 illustrates the foreground weight-based pooling strategy and presents visualizations of the confidence maps computed from optical flow gradients (f_m), color gradients (f_c) and visual saliency (f_s). We can observe that optical flow and color gradients are computed with respect

Methods	KTH	HMDB51	UCF101
BoV	91.8	24.5	43.9
FV	92.7	36.2	64.3
OFA	93.2	37.8	66.4
ForeW	92.5	38.0	65.8
Pyramid	93.0	40.2	69.2
Fusion	**93.6 (+0.9)**	**41.5 (+5.3)**	**71.8 (+7.5)**

Table 2: Average accuracies (%) of optical flow orientation-based (OFA) and foreground weight-based (ForeW) pooling strategies compared with BoV, FV, and spatio-temporal pyramid (Pyramid) method on the KTH, HMDB51, and UCF101 datasets. The average accuracies of decision-level fusion of OFA, ForeW and Pyramid are also presented. Improvements compared with the FV method are shown in the parentheses.

to motion and spatial edges, while visual saliency providing complementary information about the areas capturing main focus of human attention. The outlines of the heatmap f_{conf} are thicker than that of optical flow gradients and color gradients heatmaps. Therefore, more STIPs are incorporated in the foreground to improve the robustness of our algorithm.

We project the STIPs into two different bins according to a threshold of foreground weight. Fisher Vectors are computed for these interest points respectively. And then same pooling and normalization procedures are conducted as the optical flow orientation-based approach. These two bins are concatenated to form the final representation with a dimension of 129024.

3.1.4 *Spatio-Temporal Pyramid-based Pooling Strategy*

We also implement spatio-temporal pyramid approach with Fisher Vector encoding procedure on all of three datasets. Six spatio-temporal grids distributions are employed: $1 \times 1 \times 1$, $1 \times 3 \times 1$, $2 \times 2 \times 1$, $1 \times 1 \times 2$, $1 \times 3 \times 2$ and $2 \times 2 \times 2$. A visualization of our spatio-pyramid approach is presented in Figure 6. How to divide grids distributions to achieve best results is an engineering trick depending on different datasets. Since our paper focuses on a generalized pooling approach, we apply the same parameters of pooling strategies to all the datasets.

By decision-level fusion of spatio-temporal pyramid, optical flow orientation-based and foreground weight-based pooling methods, an improved performance is achieved compared with the baseline.

3.2 Experimental Results and Analysis

As shown in Table 2, the Fisher Vector encoding procedure outperforms BOV words approach on all datasets. The reason is that Fisher Vector is based on a generative model which encodes more higher order information, while BOV words method is based on hard assignments to generated codewords [18]. In following analysis, we choose the results of Fisher Vector encoding procedure with standard sum pooling strategy as the baseline.

The performance of the foreground weight-based pooling strategy is slightly lower than the baseline on the KTH dataset, since the background is relatively clean for the KTH dataset while the majority of the STIPs lay on the foreground scenes, which makes the *static* bin relatively redundant. For the other two datasets, both optical flow orientation-based and foreground weight-based methods have positive influence on the recognition performance. Specifically, for the HMDB51 dataset, OFA method improves performance from 36.2% to 37.8% and ForeW method improves to 38.0%. For the dataset UCF101, OFA achieves an improvement of 2.1% compared with the baseline outcome 64.3% and ForeW method has 1.5% increase.

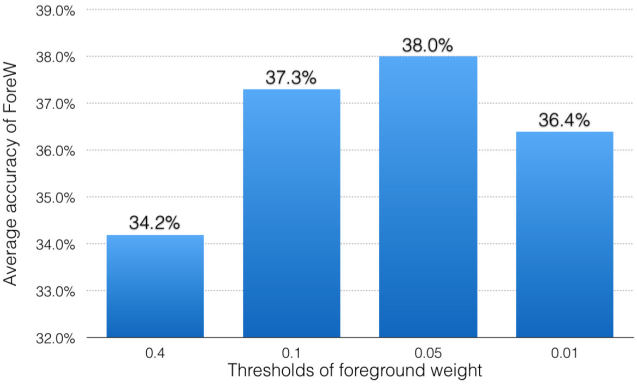

Figure 7: The average accuracies of foreground weight-based pooling strategy on the HMDB51 dataset according to different foreground thresholds.

However, except for the KTH dataset, the spatio-temporal pyramid approach consistently outperforms other methods. It improves 4.0% and 4.9% respectively for the HMDB51 dataset and the UCF101 dataset. Since the dimension of the representation of spatio-temporal pyramid approach is $24 \times 64512 = 1548288$, it is an unwise idea to fuse spatio-temporal method with the other two in the descriptor level or representation level [10]. Decision-level fusion approach is applied to produce a better recognition performance. In our experiments, we apply a geometric mean approach to the KTH and UCF101 datasets and an arithmetical mean method to the HMDB51 dataset to achieve a better performance.

As shown in the last row of Table 2, for the KTH dataset there is 0.9% boost after fusion compared with the baseline results. As for the HMDB51 and UCF101 datasets, the improvements are 5.3% and 7.5% respectively.

We also explore the influence of the selection of threshold of foreground weight on the HMDB51 dataset. As shown in Figure 7, the best result is achieved with 0.05 threshold among four thresholds: 0.4, 0.1, 0.05, 0.01. All the results of ForeW method presented in the Table 2 are obtained with the threshold 0.05.

4. CONCLUSION

In this paper, we have explored the importance of pooling strategies based on the idiosyncrasies of individual STIPs. Specifically three pooling approaches are discussed: spatio-temporal pyramid, optical flow orientation-based and foreground weight-based methods.

Unlike the commonly used sum pooling and max pooling strategies which are performed on the elements of vectors respectively in the global domain and may neglect the discriminative power of different descriptors of STIPs, our proposed pooling framework emphasizes the distinct properties of individual STIPs and therefore renders more discernability to video representations. For the local space-temporal interest points-based action recognition algorithms, the appropriate pooling strategies based on the peculiar idiosyncrasies of individual interest points can provide important complementary information to improve the final performance.

5. ACKNOWLEDGEMENT

This work was supported in part by NSF Grants IIS-1400802 and EFRI-1137172.

6. REFERENCES

[1] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, 2011.

[2] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *Computer Vision and Pattern Recognition*, 2008.

[3] Y.-L. Boureau, J. Ponce, and Y. LeCun. A theoretical analysis of feature pooling in visual recognition. In *International Conference on Machine Learning*, 2010.

[4] L. Cao, Z. Liu, and T. S. Huang. Cross-dataset action detection. In *Computer Vision and Pattern Recognition*, 2010.

[5] B. Chakraborty, M. B. Holte, T. B. Moeslund, and J. Gonzàlez. Selective spatio-temporal interest points. *Computer Vision and Image Understanding*, 2012.

[6] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005.

[7] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *International Conference on Computer Vision*, 2005.

[8] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, 2006.

[9] B. K. Horn and B. G. Schunck. Determining optical flow. In *Technical Symposium East*, 1981.

[10] N. Ikizler-Cinbis and S. Sclaroff. Object, scene and actions: Combining multiple features for human action recognition. In *European Conference on Computer Vision*. 2010.

[11] T. Jaakkola, D. Haussler, et al. Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*, 1999.

[12] A. Kläser and M. Marszalek. A spatio-temporal descriptor based on 3d-gradients. 2008.

[13] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *International Conference on Computer Vision*, 2011.

[14] I. Laptev and T. Lindeberg. Space-time interest points. In *International Conference on Computer Vision*, 2003.

[15] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition*, 2008.

[16] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition*, 2006.

[17] D. Lin, C. Lu, R. Liao, and J. Jia. Learning important spatial pooling regions for scene classification. In *Computer Vision and Pattern Recognition*, 2014.

[18] L. Liu, L. Wang, and X. Liu. In defense of soft-assignment coding. In *International Conference on Computer Vision*, 2011.

[19] B. D. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. In *International Joint Conferences on Artificial Intelligence*, 1981.

[20] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.

[21] X. Peng, Y. Qiao, Q. Peng, and X. Qi. Exploring motion boundary based sampling and spatial-temporal context descriptors for action recognition. In *British Machine Vision Conference*, 2013.

[22] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *European Conference on Computer Vision*. 2010.

[23] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *International Conference on Pattern Recognition*, 2004.

[24] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th international conference on Multimedia*, 2007.

[25] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[26] W. Sultani and I. Saleemi. Human action recognition across datasets by foreground-weighted histogram. In *Computer Vision and Pattern Recognition*, 2014.

[27] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li. Hierarchical spatio-temporal context modeling for action recognition. In *Computer Vision and Pattern Recognition*, 2009.

[28] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition*, 2011.

[29] H. Wang and C. Schmid. Action recognition with improved trajectories. In *International Conference on Computer Vision*, 2013.

[30] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid, et al. Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conference*, 2009.

[31] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *European Conference on Computer Vision*. 2008.

[32] X. Yang and Y. Tian. Action recognition using super sparse coding vector with spatio-temporal awareness. In *European Conference on Computer Vision*. 2014.