Action Recognition via Local Descriptors and Holistic Features

Xinghua Sun
Nanjing University of Science and Technology
Nanjing 210094 China
xinghuasun@mail.njust.edu.cn

Mingyu Chen Alexander Hauptmann
Carnegie Mellon University
Pittsburgh PA 15213 USA
{mychen,alex}@cs.cmu.edu

Abstract

In this paper we propose a unified action recognition framework fusing local descriptors and holistic features. The motivation is that the local descriptors and holistic features emphasize different aspects of actions and are suitable for the different types of action databases. The proposed unified framework is based on frame differencing, bag-of-words and feature fusion. We extract two kinds of local descriptors, i.e. 2D and 3D SIFT feature descriptors, both based on 2D SIFT interest points. We apply Zernike moments to extract two kinds of holistic features, one is based on single frames and the other is based on motion energy image. We perform action recognition experiments on the KTH and Weizmann databases, using Support Vector Machines. We apply the leave-one-out and pseudo leave-N-out setups, and compare our proposed approach with state-of-the-art results. Experiments show that our proposed approach is effective. Compared with other approaches our approach is more robust, more versatile, easier to compute and simpler to understand.

1. Introduction

Acton recognition has been widely researched and applied in many domains, such as visual surveillance, human computer interaction and video retrieval etc. Aggarwal and Cai [1] give an overview of the various tasks involved in the motion analysis of human body. Hu *et al.* [2] review the visual surveillance in dynamic scenes and analyze possible research directions. Generally speaking, action recognition framework contains three main steps namely feature extraction, dimension reduction and pattern classification. The feature extraction can be broadly divided into two categories, one is based on local descriptors [3-18] and the other is based on holistic features [7, 16, 19-31]. As to the dimension reduction approaches, there are PCA [3, 26], LDA [7], LLE [32], LPP [23, 31] and LSTDE [29] etc. The pattern classifier can be divided into two categories, one is based on the stochastic model such as HMM [33] and pLSA [12, 17], etc., and the other is based on the statistical model such as ANN [7, 31], NNC[3, 12], SVM [4, 8-10, 12-15, 26], LPBoost [10] and AdaBoost [25] etc.

There are several available action databases, among which the Weizmann database (see Figure 1(a)) [28] and the KTH database (see Figure 2(a)) [4] have been widely used to evaluate action recognition approaches and many results have been reported on them (see Tables 4 and 5). Some approaches are evaluated on both (e.g. [13, 15, 25, 26]) while others either only on the Weizmann (e.g. [27-31, 34]) or the KTH database (e.g. [7-12, 14, 24]). Most existing action descriptors can be divided into two categories of local descriptors [3-17] and holistic features [7, 16, 19-31]. However, some approaches do not neatly fall into these categories, e.g. Ali *et al.* [34] is based on the theory of chaotic systems.

Table 5 shows that action recognition rates, on the Weizmann database for the top existing approaches (above 90%) are not based on local descriptors, except [15] which is based on a biologically-motivated system, but most are based on holistic features [25-31]. Our proposed local descriptor approach (see row "*2D + 3D SIFT*" in Table 5) only gives a recognition rate of 90.3%, which is less than the result of our holistic feature approach of 94.6% (see row "*FRM ZNK + MEI ZNK*" in Table 5). That is to say, on the Weizmann database holistic feature approaches are superior to local descriptor approaches. In Table 4, among the results above 90% on the KTH database, five are based on local descriptors [7-9, 13, 15], three are based on holistic features [7, 25, 26] and one [7] is based on both. It is notable that the approaches in [25, 26] are based on human centered alignment. On the KTH database our proposed local descriptor approach gives an accuracy of 91.4% (see row "*2D + 3D SIFT*" in Table 4), however the result of our holistic feature approach is only 87.7% (see row "*FRM ZNK + MEI ZNK*" in Table 4). So it can be said that local descriptor approaches work very well on the KTH set and are slightly better than holistic feature approaches.

Why does the same approach have the different performance on the different database? Or, why are local descriptors more suitable for KTH, while the holistic features are more suitable for Weizmann? The answer lies in the different characteristics of these two databases. KTH has a larger data scale, four different scenarios, changing backgrounds due to the camera zoom, more persons

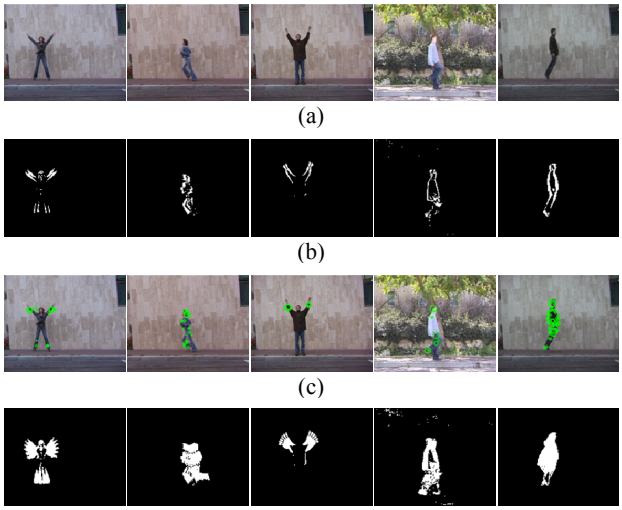


Figure 1: The Weizmann database

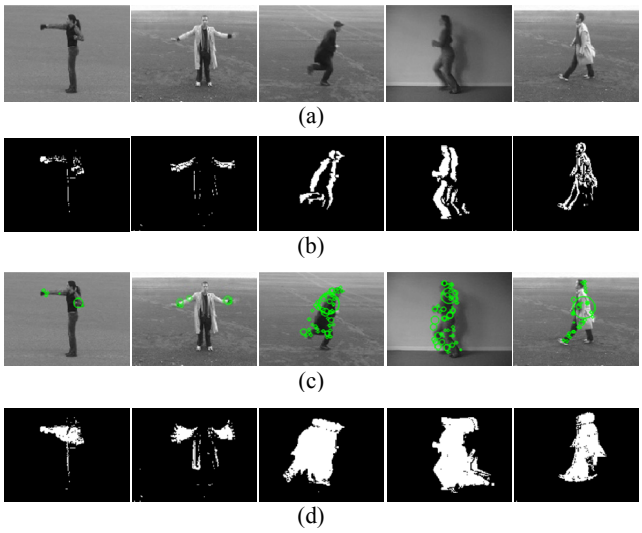


Figure 2: The KTH database

performing actions, and more intra-class dissimilarity in the shape of figures. In contrast, the Weizmann data has a much lower data scale, only one scenario, static background, more action classes and more inter-class similarity in the local motion, e.g. the jump and skip actions are very similar to each other. Local descriptor approaches extract the neighborhood information of interest points and focus more on local motion than on the figure shape. But figure shape is an advantage of holistic features as they focus more on global information. That may be the reason why local descriptor approaches deal with KTH very well while they cannot deal with Weizmann as well, even though KTH appears more challenging.

Since local descriptors and holistic features emphasize different aspects of actions and are suitable for the different databases, one natural idea is to combine them to improve the performance. Obviously this is a multiple information fusion problem. There have been some similar efforts in the action recognition field. Liu *et al.* [16] fuse multiple features for improved action recognition in videos, with a local descriptor feature and one about spin images. Mikolajczyk and Uemura [7] use local descriptors and optical flow information to form a vocabulary forest of local motion-appearance features to recognize actions. Compared with [7, 16], we apply different local descriptors and different holistic features to perform feature fusion and get better performance on both the KTH and Weizmann databases (see Table 4 and 5). Actually we use two kinds of local descriptors and two kinds of holistic features and the final fusion is based on the four sets of features. Experiments show that the feature fusion does improve the action recognition performance.

2. Related work

The main idea of this paper is to fuse local descriptors and holistic features to perform action recognition, so we review representative papers on these two different features respectively.

2.1. Local descriptors

Dollar *et al.* [3] use sparse spatiotemporal features to perform behavior recognition including human and rodent behavior. Schuldt *et al.* [4] construct video representations in terms of local space-time features and integrate such representations with SVM classification schemes for recognition. Laptev and Lindeberg [5] build on the idea of the Harris and Forstner interest point operators and detect local structures in space-time. Shechtman and Irani [6] extend the notion of 2-dimensional image correlation into a 3-dimensional space-time volume, thus enabling them to correlate dynamic behaviors and actions. Liu and Shah [8] use the Maximization of Mutual Information (MMI) technique to select the optimal number of words for bag-of-words algorithm. Laptev *et al.* [9] address recognition of natural human actions in diverse and realistic video settings. Klaser *et al.* [13] present a local descriptor based on histograms of oriented 3D spatio-temporal gradients. Wong and Cipolla [12] utilize the global information to yield a sparser set of interest points for motion recognition. Willems *et al.* [14] present the spatio-temporal interest points that are at the same time scale-invariant (both spatially and temporally). Oikonomopoulos *et al.* [18] detect the spatiotemporal salient points by measuring changes in the information content of pixel neighborhoods not only in space but also in time.

2.2. Holistic features

Bobick and Davis [21] use temporal templates, including motion-energy images and motion-history images to recognize human movement. Gorelick *et al.* [22] exploit a solution to the Poisson equation to extract various shape properties from images. Wang and Suter [23] learn explicit representations for dynamic shape manifolds of moving humans. Jia and Yeung [29] use a dimensionality reduction approach called LSTDE to recognize silhouette-based human action. Gorelick *et al.* [28] regard human actions as three dimensional shapes induced by silhouettes in the space time volume. Weinland *et al.* [19] use learned 3D exemplars to produce 2D image information to perform view-independent action recognition. Rodriguez *et al.* [24] use a frequency domain technique, called the Maximum Average Correlation Height (MACH) filter, to recognize single-cycle human actions. Wang *et al.* [31] use the Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT) to describe silhouette-based image.

3. Action recognition framework

We extract local descriptors and holistic features based on the frame differencing, and then use a bag-of-words approach to compute feature vectors for the feature fusion. Our approach may be the first to use frame differencing to extract both local descriptors and holistic features. In the previous work, most local descriptors approaches use the spatiotemporal gradient information to extract interest points and most holistic features are based on the silhouettes or tracking. In contrast, we present a unified action recognition framework for local descriptors, holistic features and their fusion.

3.1. Image segmentation

A key problem is the appropriate image representation to compute the image or video features for action recognition. The role of image segmentation is to get such an image representation and to give prominence to the object information interesting to users. In video with static background, like the Weizmann data, the silhouette [22, 28-32, 34] is a good image representation, which can be obtained by background subtraction. But for video without static background, like KTH, the background subtraction technique doesn't work. In this case one feasible option is to use the difference image between adjacent frames, which has been used to extract temporal templates by Bobick and Davis [21]. In some approaches based on local descriptors [4, 5], the lack of a static background can be solved by localizing the interest points with spatiotemporal gradient information from the original video, which cannot be done for holistic features.

Our goal in image segmentation is to find a relatively common image representation which can be used either with or without static background, both for local descriptors and for holistic features. The solution is to use the difference video, which is made up of all the difference images between adjacent frames, as illustrated in Figure 1(b) for Weizmann and in Figure 2(b) for KTH. The difference video just captures motion information and the silhouette contains both still shape and dynamic motion information. But it is difficult to judge which representation is superior over another. For example, for the two actions of waving one hand and waving two hands, the most discriminative part of silhouette should be around the arms while the torso could be considered noise. The main advantage of difference video is that it can be applied without requiring a static background. Of course, the difference video approach cannot be applied to video with dramatic background change (e.g. moving camera).

3.2. Bag of words

The bag-of-words approach has been widely used and the simplest output is a histogram reflecting the distribution of all the words. The central part is a sequence of predefined words, denoted as $\{w_j\}(1 \le j \le K)$, where K is the number of words. Here each word is just a feature vector. For each feature vector d we compute its distance to each predefined word and get the minimum distance. Then the feature vector d is assigned to the word w_{j_d} having the minimum distance to d, i.e. $j_d = \arg\min_{1 \le j \le K} \mathcal{D}(d, w_j)$. Here $\mathcal{D}(*)$ is a distance function. After all feature vectors have been assigned to words, we get a histogram $H = \{h_j\}(1 \le j \le K)$ whose bin represents how many feature vectors each word contains, i.e. the histogram H is the output of bag-of-words and can then be concatenated with other histograms in the fusion of multiple features, or just used directly as the input to classifiers to perform the action recognition.

The motivation for applying a bag-of-words approach to action recognition was to deal with the variable number of interest points produced by the local descriptors for different videos. Of course the holistic feature doesn't have an interest point related problem, but it has the problem of being sensitive to varying action duration time. The bag-of-words aggregates the statistical temporal information of a video event and therefore can deal with long-term or multiple-cycle action video. Given a video $V = \{f_i\}(1 \le i \le N)$, holistic features are computed on each frame or its variant frame (e.g. MEI, see Section 5), resulting in a sequence of feature vectors $D = \{d_i\}(1 \le i \le N)$, where f is a single frame (or its MEI), d is the holistic feature, e.g. Zernike moments, which forms the input to the bag-of-words.

3.3. Feature fusion

The basic idea of feature fusion is to concatenate all the

feature vectors produced by different approaches to form a larger feature vector as input to a classifier such as Support Vector Machine (SVM). The prerequisite to effective feature fusion is that each individual feature vector has the same physical meaning, which is guaranteed with the bag-of-words technique here. Given M approaches to action recognition based on bag-of-words, there is a sequence of feature vectors $\{H^l | H^l = \{h_j^l\}(1 \le j \le K^l)\}$, where $1 \le l \le M$. Then the larger feature vector can be denoted as $H^{FUSION} = \{h_1^1, h_2^1, \dots, h_{K^1}^1, \dots, h_1^M, h_2^M, \dots, h_{K^M}^M\}$. It should be noted that each feature vector's dimension K^l is not necessarily the same in each approach.

4. Local descriptors

We use frame differencing to localize interest points and extract local descriptors. Considering that the SIFT features have very desirable characteristics such as invariance to transformation, rotation and scale, and robustness to partial occlusion etc., we use 2D SIFT interest points to extract 2D and 3D SIFT feature descriptors. The interest points are localized based on each frame differencing, so the resulting interest points are quite dense. The 2D and 3D SIFT feature descriptors emphasize the still shape and dynamic motion respectively. Note that our local descriptors only provide spatial scale invariance, not the temporal and spatial scale invariance demonstrated in [14].

4.1. 2D SIFT interest point

To be scale invariant, SIFT features need to consider all the scales of an image. Lowe [35] uses the Gaussian function as the scale-space kernel to produce the scale space of the image. Given an image $I(r,c)$ the scale space is produced as $L(r,c,\sigma) = G(r,c,\sigma) * I(r,c)$ where $*$ is the convolution operation in r and c, and $G(r,c,\sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{r^2+c^2}{2\sigma^2}}$. The whole scale space is divided into a sequence of octaves and each octave is divided into a sequence of intervals. Here each interval is a scale image. The relationship between two adjacent octaves is that the first interval in the latter octave is gotten by down-sampling the last interval in the former octave by a factor of 2. The first interval in the first octave is just $I(r,c)$. This yields a pyramid-like structure of Gaussian space.

After the pyramid-like structure of scale space is created, in each octave every interval is subtracted from its adjacent interval to compute the difference-of-Gaussian (DOG) function, i.e. $D(r,c,\sigma) = L(r,c,k\sigma) - L(r,c,\sigma)$. Thus we get a pyramid-like structure of DOG space. Given a pixel in the DOG space, the DOG intensity is compared with its eight neighbors in the same interval and its nine neighbors in the interval above and below. If the DOG intensity is larger than all of these neighbors or smaller than all of them, the pixel is selected as an interest point. Figure 1(c) and

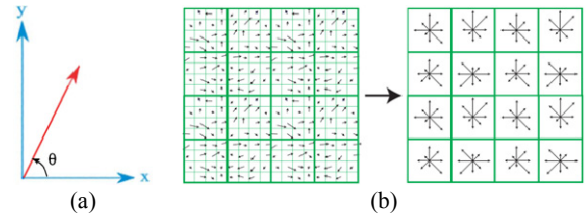


Figure 3: Computation of the 2D SIFT feature descriptor

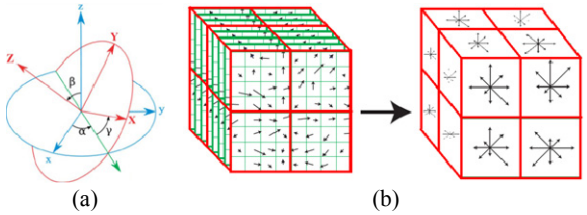


Figure 4: Computation of the 3D SIFT feature descriptor

Figure 2(c) give examples of interest points for Weizmann and KTH respectively. We use a green circle to represent an interest point with the center reflecting the position and the radius reflecting the gradient magnitude.

4.2. 2D SIFT feature descriptor

The SIFT feature descriptor is extracted based on Gaussian space. The gradient magnitude $m(r,c)$ and orientation $\theta(r,c)$ (see Figure 3(a)) is computed at each scale image. For an interest point, the gradient magnitude and orientation of all the pixels in its neighborhood are used to produce an orientation histogram, where the magnitude is used as a weight to calculate each bin. With the orientation histogram, dominant orientations can be detected in the largest bins. For each dominant orientation, the neighborhood of the interest point will be rotated to get rotation invariance. The size of the neighborhood region in [35] is set to be 16 x 16 in pixels. Then the whole region is divided into multiple sub-regions whose size is 4 x 4 in pixels and the total number of sub-regions is 16, as illustrated in Figure 3(b). In each sub-region, an orientation histogram is produced with each orientation histogram having 8 bins. All the histograms of these 16 sub-regions are combined to form a 128 (= 16 x 8) dimensional feature vector, which is just the 2D SIFT feature descriptor.

4.3. 3D SIFT feature descriptor

Similar to 2D, the 3D SIFT feature descriptor is based on Gaussian space. Here each interest point has five dimensions. The gradient information in the video contains one magnitude ($m(f,r,c)$) and three orientations. These three orientations reflect the angles between row and column ($\alpha(f,r,c)$), between row and time ($\beta(f,r,c)$), between column and time ($\gamma(f,r,c)$), respectively, as

illustrated in Figure 4(a). According to the neighborhood gradient information, the orientation histogram is produced and the dominant orientations are detected. For each dominant orientation the neighborhood of interest point is rotated to get rotation invariance. Here the neighborhood is a cube, whose size is set to be 8 x 8 x 8 in pixels. The whole cube will be divided into multiple sub-cubes whose size is 4 x 4 x 4 in pixels and the total number of sub-cubes is 8, as illustrated in Figure 4(b). For each sub-cube and each orientation, the orientation histogram with 8 bins is produced. Thus we have 24 orientation histograms in the whole cube. All these histograms are combined to form a 192 (= 24 x 8) dimensional feature vector, which is the 3D SIFT feature descriptor.

5. Holistic features

We use the frame differencing to compute holistic features, avoiding object tracking (which is often not reliable) and silhouette extraction (which needs a static background). We compute two kinds of holistic features with Zernike moments, one is based on a single fame and the other is based on the image motion energy, where the former is focused on spatial patterns and the latter is focused on temporal constraints.

5.1. Zernike moments

Hu moments [36] have been widely used in pattern recognition as holistic features. Prokop and Reeves [37] proved that Zernike moments are best among multiple invariant moments in terms of overall performance. So in this paper we use Zernike moments as holistic features to perform the action recognition instead of Hu moments. As the essential action information, motion is depicted with temporal templates including MHI and MEI in [21]. The bag-of-words technique is also an effective tool for representing motion information, although it just captures implicit motion information. Thus, the holistic features of a single frame can be used in the bag-of-words framework. To deal with the long-term or multiple-cycle action, the temporal granularity for the computation of holistic features should be less than one action cycle. The extreme case is to choose a single frame as the computational granularity. The combination of bag-of-words with Zernike moments based on single frame is used as the holistic approach component to our action recognition, called FRM ZNK.

5.2. Motion energy image

The shortcoming of the bag-of-words approach for describing motion is that the temporal constraint information between frames cannot be preserved. Of course, this shortcoming is an inherent characteristics of histogram techniques, they just provide statistical temporal information, which is the reason they are robust to noise. To complement the bag-of-words based on the holistic features of a single frame, we also extract the Motion Energy Image (MEI, illustrated in Figure 1(d) for Weizmann and in Figure 2(d) for KTH) and compute the corresponding Zernike moments as holistic features for the bag-of-words. We call this part MEI ZNK.

The reasoning behind MEI ZNK is that the MEI can provide temporal constraint information. One advantage of MEI is that it can provide robustness to some kinds of noise motion. For example, depending on the texture of the people, several "inside" regions may or may not produce motion (consider a person wearing a black shirt as compared to a striped shirt). MEI aggregates the motion information across several frames and the above noise motion produced by "inside" regions will be overlaid by the torso motion. The time duration for extracting MEI is critical and must be less than a single action cycle. Empirical observations show that a single action cycle in KTH and Weizmann can be as short as 5 frames. So in experiments the duration is set to 5. FRM ZNK and MEI ZNK together constitute our holistic approach to action recognition.

6. Experiments

We use KTH and Weizmann to evaluate action recognition. We perform leave-one-out experiments and the result is reported as the average. K-means clustering is applied to the descriptor bag-of-words to get the predefined words. In both databases, we use the data of 2 persons to perform clustering. We use Support Vector Machines (SVM) as the classifier and try two kinds of kernels, the polynomial kernel (POLY SVM) and the radial basis function kernel (RBF SVM). With the aid of LibSVM [38], the parameters for RBF kernel is optimized and the normalization of features is needed.

For the bag-of-words the number of words is a key parameter. We use POLY SVM to test 10 numbers for both databases, as illustrated in Figure 5. The results for the optimal number of words are listed in Table 1. It can be seen that the optimal number of words depends on two aspects, the type of feature and the scale of the database. With the same feature, a larger database means a larger optimal number of words. We use these configurations of optimal numbers of words for all other experiments, including RBF SVM and feature fusion.

	Local		Holistic	
	2D SIFT	3D SIFT	FRM ZNK	MEI ZNK
KTH	500	400	200	500
Weizmann	60	50	30	100

Table 1. Optimal number of words

Tables 2 and 3 give the feature fusion results on KTH and Weizmann data respectively. The performance after fusing two local descriptors is larger than the maximum individual by about 3% on KTH (Table 2(a)) and 12% on Weizmann (Table 3(a)). The fusion of holistic features can bring a performance improvement of about 3% on KTH (Table 2(b)) and 1% on Weizmann (Table 3(b)). The best action recognition rate results from the fusion of local descriptors and holistic features, i.e. 94% on KTH (Table 2(c)) and 97.8% on Weizmann (Table 3(c)), both of which are larger than the maximum individual by about 3%. These two tables clearly show that fusing different categories of features, such as local descriptors and holistic features **does** improve action recognition performance.

	2D SIFT	3D SIFT	2D + 3D SIFT
POLY	83.1	86.8	90.5
RBF	86.3	88.6	**91.4**

(a)

	FRM ZNK	MEI ZNK	FRM ZNK + MEI ZNK
POLY	84.7	82.6	85.0
RBF	85.1	83.5	**87.7**

(b)

	SIFT	ZNK	SIFT + ZNK
POLY	90.5	85.0	89.5
RBF	91.4	87.7	**94.0**

(c)

Table 2. Feature fusion results on KTH data

	2D SIFT	3D SIFT	2D + 3D SIFT
POLY	78.5	69.9	75.3
RBF	78.5	76.3	**90.3**

(a)

	FRM ZNK	MEI ZNK	FRM ZNK + MEI ZNK
POLY	83.9	87.1	88.2
RBF	87.1	93.5	**94.6**

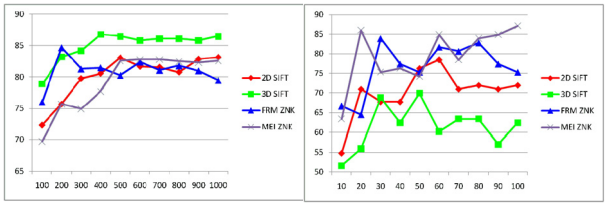
(b)

	SIFT	ZNK	SIFT + ZNK
POLY	75.3	88.2	92.5
RBF	90.3	94.6	**97.8**

(c)

Table 3. Feature fusion results on Weizmann data

On KTH data SIFT features result in better performance than Zernike moments, i.e. 91.4% vs. 87.7%. On Weizmann data the Zernike moments are superior to SIFT features with a recognition rate of 94.6% vs. 90.3%. From this observation, we conclude that no one single category of



(a) (b)

Figure 5: Action recognition rates with different numbers of words for bag-of-words, where (a) is for KTH and (b) is for Weizmann. The peak of each line means the optimal number of words.

(a) (b)

Figure 6: Confusion matrixes, where (a) is for KTH and (b) is for Weizmann

feature can deal with all kinds of action databases equally well. So it is quite necessary and useful to fuse different categories of features to improve the action recognition performance. Figure 6 gives the confusion matrices of our best results on (a) KTH and (b) Weizmann data. Instead of simple concatenation, we also try a weighted concatenation in the feature fusion. The weights are assigned according to the recognition rate of individual feature (higher rate, higher weight) and the dimension of histogram (larger dimension, lower weight). The resulting recognition rate on KTH is 93.3%, without any improvement as compared to the simple concatenation.

Tables 4 and 5 compare our action recognition result with state-of-the-art results on KTH and Weizmann data respectively. The basic description of the compared approaches can be found in Section 2. On KTH, our final recognition rate is 94.0%, just barely less than the current best rate by 0.2%. Due to the relatively small amount of data, the performance on Weizmann data tops out at 100% and our rate of 97.8% is very close to that. It should be noted that different approaches have different experimental configurations. Some approaches divide the database into two parts for training and testing respectively [4, 9-11, 13-15, 25, 26] instead of leave-one-out. Some use human tracking techniques to align the image with a human center [25-27, 30] and others use supervised training approaches. Our approach is focused on an automatic training and testing process without any human involvement, e.g. no normalization is taken to make all sequence segments have the same length. The bag-of-words model is used for the representation of the whole sequence without any feature selection.

Approaches	Rates (%)
SIFT + ZNK	**94.0**
2D + 3D SIFT	**91.4**
FRM ZNK + MEI ZNK	**87.7**
Liu and Shah [8]	94.2
Mikolajczyk and Uemura [7]	93.2
Schindler and Gool [26]	92.7
Laptev *et al.* [9]	91.8
Jhuang *et al.* [15]	91.7
Klaser *et al.* [13]	91.4
Fathi and Mori [25]	90.5
Gilbert *et al.* [11]	89.9
Rodriguez *et al.* [24]	88.7
Nowozin *et al.* [10]	87.0
Wong and Cipolla [12]	86.6
Willems *et al.* [14]	84.3
Niebles *et al.* [17]	81.5
Dollar *et al.* [3]	81.2

Table 4. Performance comparison on KTH

Approaches	Rates (%)
SIFT + ZNK	**97.8**
2D + 3D SIFT	**90.3**
FRM ZNK + MEI ZNK	**94.6**
Schindler and Gool [26]	100.0
Fathi and Mori [25]	100.0
Weinland and Boyer [30]	100.0
Gorelick *et al.* [28]	99.6
Jhuang *et al.* [15]	98.8
Wang *et al.* [31]	97.8
Thurau and Hlavac [27]	94.4
Ali *et al.* [34]	92.6
Jia and Yeung [29]	90.9
Liu *et al.* [16]	89.3
Klaser *et al.* [13]	84.3

Table 5. Performance comparison on Weizmann

Considering that the leave-one-out experimental setup is easier than the standard setup based on the training/test split of the database used originally in [4], we give the results based on a pseudo leave-N-out setup (see Table 6). For example, in the KTH database 25 persons form a loop. Each person and sequential N-1 persons in the loop are used for testing with the remainder for training. So the experiment will be performed by 25 times and the result is reported as the average (AVG), maximum (MAX) and minimum (MIN) of 25 runs. In Table 6, the first line means the different N for pseudo leave-N-out, (a) is for KTH and (b) is for Weizmann. From the gap between MAX and MIN in Table 6, it can be seen that the different selection of training and testing sets often lead to the different recognition rate. From this viewpoint the pseudo leave-N-out setup is more reasonable than the standard setup used in [4].

To compare with other approaches that split the datasets, we also perform the experiment based on the standard setup in [4] (with the rate of 89.8%) and the experiment based on the pseudo leave-9-out setup. In the latter case the validation set (involving 8 persons) in [4] is used to optimize the SVM models and the residual 17 persons are used to perform the pseudo leave-9-out experiment. The result is reported as the average (AVG = 89.1%), maximum (MAX = 93.5%) and minimum (MIN = 85.5%) of 17 runs, each of which is just one experiment based on the standard setup in [4]. Experiments show that our proposed approach is effective. Compared with other approaches our approach is more robust, more versatile, easier to compute and simpler to understand.

	1	5	10	15	20
AVG	94.0	92.1	91.0	88.7	81.3
MAX	99.0	96.5	94.1	90.8	87.3
MIN	78.9	87.5	88.8	86.1	71.9

(a)

	1	3	6
AVG	97.8	94.6	81.0
MAX	100.0	100.0	87.3
MIN	90.0	90.0	74.6

(b)

Table 6. The results based from the pseudo leave-N-out setup

7. Conclusions

In this paper we reviewed existing action recognition approaches' performance on the KTH and Weizmann databases, and analyzed why local descriptors seem more suitable for KTH while the holistic features seem more suitable for Weizmann. Based on our analysis we proposed a unified action recognition framework fusing local descriptors and holistic features. Experiments show that our proposed approach is effective and its final performance is comparable to other published results. The fusion approach adopted in this paper is a rather simplistic manner of fusion and in the future we will try the more flexible fusion approach and feature selection approaches to get a best subset of features. Due to the usage of difference video, our approach will fail with the dramatic background change, which often appears in movies. So we will investigate the performance of our approach in the Hollywood data presented in [9].

8. Acknowledgements

This material is based in part upon work supported by the National Science Foundation under Grants No. 0624236 and 0751185. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

The most part of this work is finished in CMU, where Xinghua Sun is hosted as a visiting scholar by Alexander Hauptmann. Xinghua Sun is partially supported by the NSFC (No. 60705020), the Science Foundation of Jiangsu Province (No. BK2007594), and the high-tech plan of Jiangsu Province (No. BG2005008).

References

[1] J. K. Aggarwal and Q. Cai. Human motion analysis: a review. *IEEE Proceedings of Nonrigid and Articulated Motion Workshop*, pages 90-102, 1997.

[2] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *SMC, Part C: Applications and Reviews,* 34(3):334-352, 2004.

[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65- 72, 2005.

[4] Schuldt, C. Laptev, and B. I. Caputo. Recognizing human actions: a local SVM approach. *ICPR(17)*, pages 32-36, 2004.

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

[6] E. Shechtman and M. Irani. Space-time behavior-based correlation - OR - How to tell if two underlying motion fields are similar without computing them?. *PAMI,* 29(11):2045-2056, 2007.

[7] K. Mikolajczyk and H. Uemura. Action recognition with motion-appearance vocabulary forest. *CVPR*, pages 1-8, 2008.

[8] J. Liu and M. Shah. Learning human actions via information maximization. *CVPR*, pages 1-8, 2008.

[9] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. *CVPR*, pages 1-8, 2008.

[10] S. Nowozin, G. o. Bakır, and K. Tsuda. Discriminative subsequence mining for action classification. *ICCV*, pages 1-8, 2007.

[11] A. Gilbert, J. Illingworth, and R. Bowden. Scale invariant action recognition using compound features mined from dense spatio-temporal corners. *ECCV*, pages 222-233, 2008.

[12] S.-F. Wong and R. Cipolla. Extracting spatiotemporal interest points using global information. *ICCV*, pages 1-8, 2007.

[13] A. Klaser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. *BMVC*, pages xx-yy, 2008.

[14] G. Willems, T. Tuytelaars, and L. V. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. *ECCV*, pages 650-663, 2008.

[15] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. *ICCV*, pages 1-8, 2007.

[16] J. Liu, S. Ali, and M. Shah. Recognizing human actions using multiple features. *CVPR*, pages 1-8, 2008.

[17] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *IJCV,* 79(3):299-318, 2008.

[18] A. Oikonomopoulos, L. Patras, and M. Pantic. Spatiotemporal saliency for human action recognition. *ICME*, pages 1-4, 2005.

[19] D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3D exemplars. *ICCV*, pages 1-7, 2007.

[20] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. *ICCV(2)*, pages 726-733, 2003.

[21] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *PAMI,* 23(3):257-267, 2001.

[22] L. Gorelick, M. Galun, E. Sharon, R. Basri, and A. Brandt. Shape representation and classification using the Poisson Equation. *CVPR(2)*, pages II-61 - II-67, 2004.

[23] L. Wang and D. Suter. Learning and matching of dynamic shape manifolds for human action recognition. *IEEE Transactions on Image Processing,* 16(6):1646-1661, 2007.

[24] M. D. Rodriguez, J. Ahmed, and M. Shah. Action MACH: a spatio-temporal maximum average correlation height filter for action recognition. *CVPR*, pages 1-8, 2008.

[25] A. Fathi and G. Mori. Action recognition by learning mid-level motion features. *CVPR*, pages 1-8, 2008.

[26] K. Schindler and L. v. Gool. Action snippets: how many frames does human action recognition require?. *CVPR*, pages 1-8, 2008.

[27] C. Thurau and V. Hlavac. Pose primitive based human action recognition in videos or still images. *CVPR*, pages 1-8, 2008.

[28] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *PAMI*. 29(12):2247-2253, 2007.

[29] K. Jia and D.-Y. Yeung. Human action recognition using local spatio-temporal discriminant embedding. *CVPR*, pages 1-8, 2008.

[30] D. Weinland and E. Boyer. Action recognition using exemplar-based embedding. *CVPR*, pages 1-8, 2008.

[31] L. Wang, X. Geng, C. Leckie, and R. Kotagiri. Moving shape dynamics: a signal processing perspective. *CVPR*, pages 1-8, 2008.

[32] T. Ding. A robust identification approach to gait recognition. *CVPR*, pages 1-8, 2008.

[33] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden Markov model. *CVPR*, pages 379-385, 1992.

[34] S. Ali, A. Basharat, and M. Shah. Chaotic invariants for human action recognition. *ICCV*, pages 1-8, 2007.

[35] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV,* 60(2):91-110, 2004.

[36] M.-K. Hu. Visual pattern recognition by moment invariants. *IEEE Transactions on Information Theory,* 8(2):179-187, 1962.

[37] R. J. Prokop and A. P. Reeves. A survey of moment-based techniques for unoccluded object representation and recognition. *CVGIP Graphical models and Image Processing,* 54(5):438–460, 1992.

[38] C.-C. Chang and C.-J. Lin. LIBSVM : a library for support vector machines. 2001.