Human Action Description Based on Temporal Pyramid Histograms

Yingying Liu and Arcot Sowmya
School of Computer Science and Engineering, The University of New South Wales, Kensington, Australia

Keywords: Action Recognition, Action Description, Temporal Pyramid Histograms.

Abstract: In this paper, we present an approach to action description based on temporal pyramid histograms. Bag of features is a widely used action recognition framework based on local features, for example spatio-temporal feature points. Although it outperforms other approaches on several public datasets, sequencing information is ignored. Instead of only calculating the occurrence of code words, we also encode their temporal layout in this work. The proposed temporal pyramid histograms descriptor is a set of histogram atoms generated from the original video clip and its subsequences. To classify actions based on the temporal pyramid histograms descriptor, we design a function to calculate the weights of the histogram atoms according to the corresponding sequence lengths. We test the descriptor using nearest neighbour for classification. Experimental results show that, in comparison to the state-of-the-art, our description approach improves action recognition accuracy.

1 INTRODUCTION

Human action recognition in videos is a challenging and popular topic in computer vision with several promising applications, such as video surveillance, video indexing and human-computer interaction.

Action description is one of the key issues in human action recognition. A variety of feature extraction and description approaches have been proposed and applied to human action recognition. These approaches can be divided into two types, namely global and local representations. Global representations encode the human actions in a video as a whole, while local representations describe the human actions as a collection of local descriptors or patches (Poppe, 2010).

Global representations are obtained by stacking the features over all of the video frames. For example, (Bobick and Davis, 2001) proposed an approach based on stacking the human silhouette. They calculate 7 Hu moments of the Motion Energy Image (MEI) and Motion History Image (MHI), which are obtained from all of the video frames. There are several variations on the silhouette. For example (Wang et al., 2007) applied the R transform on the human silhouette and (Blank et al., 2005) applied Poisson equation on the stacked silhouette volume. Besides the silhouette, optical flow is also used for action representation. (Efros et al., 2003) introduced optical flow based human action recognition to sports footage. Based on their experiments, the approach can work when the human size is small and video resolution is relatively low. Although these global representation approaches have been tested effectively on several public and private datasets, there are some drawbacks. The global approaches are sensitive to noise, partial occlusion and variation in viewpoints. They require background subtraction or human tracking, which is hard to do when the video content is complex, for example, when the background is cluttered or the camera is moving, or other moving objects occur in the video.

Local representations do not require preprocessing, which makes them robust and more suitable for complex action recognition. The most popular local representations are space-time feature points, which are an extension of related 2D representations. (Laptev and Lindeberg, 2003) extended 2D Harris functions to the time dimension to obtain 3D data volume. Local maxima are selected as the 3D feature points. The approach has been found to be effective on several public datasets, including the Hollywood2 human action and scenes dataset, which consist of video clips extracted from movies (Marszalek et al., 2009). The drawback of this approach is that the number of feature points is quite low, and may not be sufficient to recognize complex action types. To solve the problem, (Dollár et al., 2005) proposed an approach based on separable filters. They applied separable Gabor filters on the video volume,

Liu Y. and Sowmya A..
Human Action Description Based on Temporal Pyramid Histograms.
DOI: 10.5220/0004825206290636
In *Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods* (ICPRAM-2014), pages 629-636
ISBN: 978-989-758-018-5

and then picked the local maxima as the feature points. Besides these methods, (Oikonomopoulos et al., 2005) extended the 2D salient point detector into XYT space, and (Willems et al., 2008) extended 2D SURF into XYT space. (Scovanner et al., 2007) introduced a 3D SIFT descriptor. (Wang et al., 2009) provide a review on different spatio-temporal features for action recognition. Bag-of-features is widely employed when using local features. Bag-of-features utilizes the statistical characteristics of feature points. After clustering the feature points into a codebook, recognition is achieved on the histograms of code words, which measures the occurrence of features.

Compared to global representations, local representations have achieved better performance on several public datasets. However, sequencing (temporal) information in videos is ignored in the bag-of-features framework, as the letter only encode the occurrence of code words. (Niebles et al., 2006) introduced an unsupervised action recognition approach based on spatio-temporal feature points and video sequential structure learned by pLSA. (Gilbert et al., 2009) learned a multi stage classifier from simple features. These two approaches mainly focus on the learning approaches. (Sun et al., 2009) proposed a hierarchical spatio-temporal context modeling approach for action recognition with point-level context (SIFT average descriptor), intra-trajectory context (trajectory transition descriptor) and inter-trajectory context (trajectory proximity descriptor). (Choi et al., 2008) introduced a spatio-temporal pyramid matching for sports videos from both dynamic features (optical flow) and static features (SIFT).

In this work, we introduce a new temporal pyramid histogram description approach by exploring the temporal structure of spatio-temporal feature points. In our approach, an action video is described by temporal pyramid histograms, which is a set of histograms of code words obtained from the original video clip and its subsequences. Therefore, the temporal pyramid histogram not only encodes the occurrence of the code words, but also their temporal layout.

1.1 Related Work

The closest related work is that due to (Bosch et al., 2007) and (Lazebnik et al., 2006).

Bosch et al. use a pyramid histogram of gradients as an object descriptor encoding object shape and spatial layout. They divide an image using increasingly finer spatial grids by repeatedly doubling the number of divisions in each axis direction, and then describe each grid by a histogram of orientation gradients (HOG). Our approach proposes a temporal pyramid histogram to describe actions and their temporal layout. An action video is segmented into subsequences of the same length repeatedly and then each of the video sequences, including the orginal sequence and the subsequences, is described by a histogram of code words.

Spatial pyramid matching for recognizing natural scene categories was introduced by Lazebnik et al. In their approach, they employ a spatial pyramid and compute histograms of local features, and weight the regions at each level as a power of 2. We employ a similar weighting scheme, except that relative weights of regions are differently set.

1.2 Contribution

The contribution of this work is the novel temporal pyramid descriptor of actions. We describe an action video by a set of histogram atoms corresponding to the original video clip and its subsequences. In our approach, spatio-temporal feature points are extracted, followed by codebook generation using kmeans clustering. Then we divide the video clip into its subsequences repeatedly, and obtain the histogram of codewords corresponding to the original video clip and its subsequences. Based on the length of a sequence, a weight is assigned to its corresponding histogram. In the end, we classify actions by a 1-NN classifier with weighted histograms. Experimental results show that our approach improves action recognition accuracy, compared to the state-of-the-art.

The rest of this paper is organized as follows. In section 2, we describe the temporal pyramid histograms descriptor in detail. In section 3, we describe the process of extracting the temporal pyramid histograms for action recognition in videos. In section 4, experimental results are shown. The summary and conclusion of our work is in section 5.

2 ACTION DESCRIPTION WITH TEMPORAL PYRAMID HISTOGRAMS

Our objective is to represent the actions in a video by both the occurrence and their temporal layout of their code words. This is based on the observation that an action consists of a series of action atoms, each of which can be described individually. For example, “hand waving” consists of “lift arms up” and “put arms down”. In this paper, rather than seperating the action atoms in each action, we divides the orig-

inal video clip into subsequences of the same length repeatedly, and describe them with a set of histograms corresponding to the subsequences, so that the representation not only encodes the code word occurrences but also their temporal variation.

2.1 Temporal Pyramid Histograms

The temporal pyramid histograms description consists of a set of histograms generated from the original video clip and its subsequences. Ideally, the subsequences should correspond to different action atoms. However, we simplify the segmentation problem by subdividing each video clip into two subsequences of the same length repeatedly, and obtain the corresponding histograms of code words. All the histograms computed on the same video are concatenated into a histogram matrix to obtain the video descriptor, written in the following form:

$$\begin{aligned} pH &= [H_1, H_2, H_3, ..., H_l, ...H_L] \\ H_l &= [h_1, h_2, h_3, ..., h_i, ...h_n] \end{aligned} \tag{1}$$

pH is the temporal pyramid histograms action descriptor; H_l are the histograms on the l^{th} layer; L is the number of temporal layers. On the l^{th} layer, the histogram set H_l consists of $n = 2^{l-1}$ histograms, which are the histograms of code words obtained from corresponding video fragments. Note that every histogram in pH is normalized.

For example, in Fig. 1, with the single action "hand waving", we set the layer number to 2. Therefore we get the temporal pyramid histograms:

$$\begin{aligned} pH &= [H_1, H_2] \\ H_1 &= [h_1] \\ H_2 &= [h_{lift \quad arms \quad up}, \quad h_{put \quad arms \quad down}] \end{aligned} \tag{2}$$

On layer 1, H_1 contains the histogram h_1 of the whole video, which is obtained from all the frames in the video. On layer 2, we divide the video into two equal subsequences, which correspond to "lift arms up" and "put arms down" respectively. Therefore H_2 contains two histogram atoms $H_{lift \quad arms \quad up}$ and $H_{put \quad arms \quad down}$.

2.2 Weight Definition

After the temporal pyramid histogram is obtained, we introduce a weight function to assign weights to the histogram atoms. In Lazebnik et al's (2006) work, a power function of 2 is employed as the weight function of the regions on each level. We too employ a power function of 2 as the weight function.

$$W = [w_1, w_2, w_3, ..., w_l, ...w_L] \tag{3}$$

where $w_l = 1/2^{l-1}$ is the weight on the l^{th} layer. The weights of the action atoms are related to the lengths of their corresponding sequences. In our approach, longer sequences have larger weights, as they are likely to be more informative about actions.

3 HUMAN ACTION RECOGNITION WITH TEMPORAL PYRAMID HISTOGRAMS

In this section, we discuss the application of the temporal pyramid histograms to human action recognition. A flowchart describing the action recognition process based on the proposed temporal pyramid histograms descriptor is shown in Fig 2. At first, we extract spatio-temporal feature points using existing algorithms. Then we calculate the temporal pyramid histograms by dividing the video sequence into subsequences repeatedly. Finally, we test the descriptor using a 1-NN classifier with weighted histograms.

3.1 Temporal Pyramid Histograms Description

In our work, we utilize the cuboid feature points proposed by (Dollár et al., 2005). Although there are several different spatio-temporal feature point extraction approaches, most of them generate sparse feature points due to rarity. That makes it difficult to employ in our algorithm. Because we also need to describe small subsequences, they might contain no feature points at all if we apply these feature points extraction algorithms. Dollár et al.'s cuboid feature points approach errs on the side of generating more feature points rather than too few. Therefore, we choose to extract the cuboid feature points, which are detected by calculating the response function based on linear separable Gabor filters:

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \tag{4}$$

where g(x,y;δ) is the 2D Gaussian smoothing kernel, applied along the spatial dimension; $h_{ev}(t;\tau,\omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2}$ and $h_{od}(t;\tau,\omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2}$ are a quadrature pair of 1D Gabor filters applied temporally. The parameters δ and τ correspond to the spatial and temporal scales of the detector. The detector has strong responses on the regions with spatially distinguishing characteristics undergoing motion. For description, Dollár et al. translated the cuboid into feature descriptors, for example gradient, windowed op-

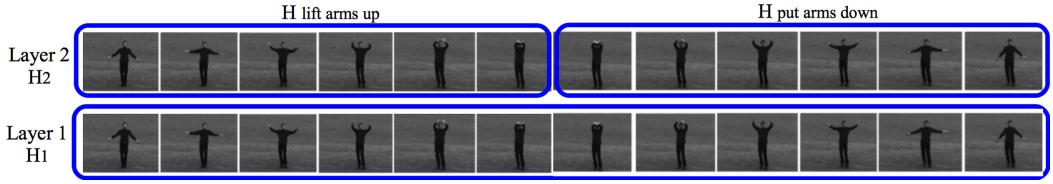

Figure 1: Illustration of temporal pyramid histogram.

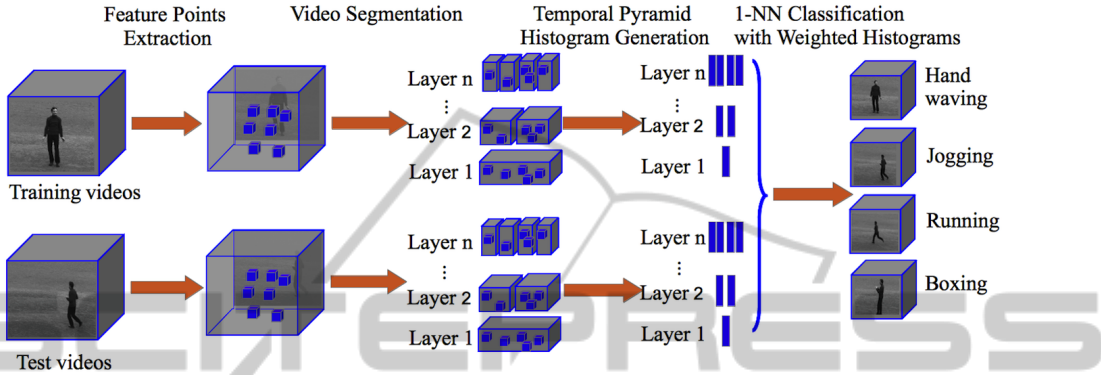

Figure 2: Flowchart of action recognition based on temporal pyramid histogram.

tical flow and normalized brightness. We use the gradient descriptor because it has the best performance according to (Dollár et al., 2005).

After obtaining the spatio-temporal feature points, videos are segmented into subsequences and the corresponding histograms computed. This is done on both training and test videos. In our approach, we segment a video sequence into subsequences repeatedly. Histograms corresponding to these subsequences are calculated, and a feature codebook generated by kmeans clustering. The action descriptor is a histogram set including all the subsequence histograms. By this method, the structure of the actions is captured in the temporal pyramid histogram.

3.2 Classification with Weighted Histograms

Similar to (Dollár et al., 2005)'s work, the 1-nearest neighbour classifier is used as the classifier in our experiments.

As described in section 2, different weights are assigned to the histograms. Therefore, based on the structure of the proposed temporal pyramid histogram descriptor, we introduce a the weight function:

$$\begin{aligned} &For \quad H^i = \{h_1^i, h_2^i, ..., h_N^i\} \\ &and \quad H^j = \{h_1^j, h_2^j, ..., h_N^j\}, \\ &K(H^i, H^j) = \sum_{k \in N} (w_k * d(h_k^i, h_k^j)) \end{aligned} \tag{5}$$

where H^i and H^j are the video descriptors of video clip i and video clip j respectively. $K(H^i, H^j)$ measures the distance between the two clips; w_k is the weight of the k^{th} histogram, which is defined in (3) according to its layer number; $d_l(h_k^i, d_k^j)$ is the distance metric between the two histograms h_k^i and h_k^j. In our

Figure 3: Sample frames corresponding to different types of facial expression (Dollár et al., 2005).

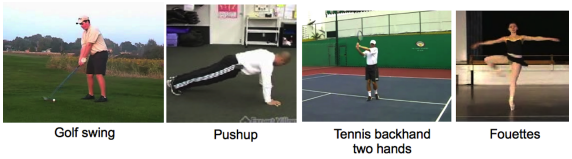

Figure 4: Sample frames corresponding to different types of action in UCF-CIL dataset (Shen and Foroosh, 2008).

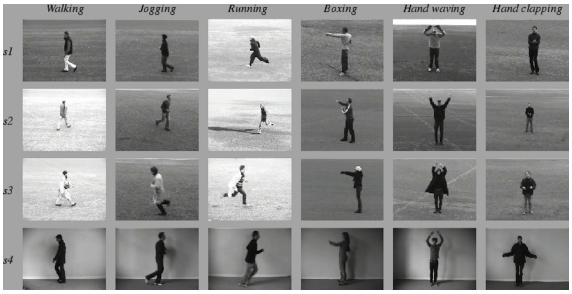

Figure 5: KTH sample frames corresponding to different types of actions and scenarios (Schuldt et al., 2004).

work, we use euclidean distance as it has the best performance in our experiments. The weight function measures the similarity between two videos. Finally, 1-NN classification is performed on the weighted histograms.

4 EXPERIMENTS

Based on the structure of the proposed pyramid temporal histograms representation method, the ideal data for the approach is well-aligned videos containing only one nonrepeating action. However, most available public datasets contain repeating actions, e.g., walking. Also videos in these datasets were not always well-aligned, because different videos of the same action were not recorded from the same pose and view. Therefore, to explore the capability of the proposed action descriptor, we tested it on a facial expression dataset (well-aligned, nonrepeating action), a subset of the UCF-CIL action dataset (not well-aligned, nonrepeating action), and KTH action dataset (not well-aligned, repeating action).

The facial expression dataset was created by (Dollár et al., 2005). There are 6 types of facial expressions (anger, disgust, fear, joy, sadness and surprise) in the dataset, performed by two individuals under two different lighting setups. Each video starts with a neutral expression, followed by an emotion, and then returns to neutral. All the videos are about 2 seconds long. There are 192 videos in total and the clips are well-aligned. Sample frames corresponding to different types of facial expressions are shown in Fig. 3.

The UCF-CIL action dataset was introduced by (Shen and Foroosh, 2008). It consists of 56 sequences of 8 actions: 4 of ballet fouettes, 12 of ballet spin, 6 of push-up exercise, 8 of golf-swing, 4 of one-handed tennis backhand stroke, 8 of two-handed tennis backhand stroke, 4 of tennis forehand stroke and 10 of tennis serve. The videos are taken from the internet. Clips in each group may have different starting and ending times. Sample frames corresponding to different types of actions in UCF-CIL dataset are shown in Fig. 4.

The KTH dataset (Schuldt et al., 2004) contains six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) performed several times by 25 subjects in four scenarios. Each video contains an action performed by one person. The database contains 2391 sequences. Each action is repeated several times in each video. Sample frames corresponding to different types of actions and scenarios are shown in Fig. 5.

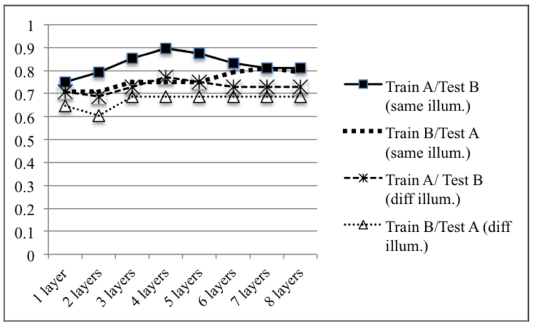

Figure 6: Average recognition accuracy on facial expression dataset. The four experiments are set up as: train on subject A and test on subject B under the same illumination; train on subject B and test on subject A under the same illumination; train on subject A and test on subject B under different illumination; train on subject B and test on subject A under different illumination.

Because of different parameter and experiment settings, we do not attempt to compare our approach with other published results. However, the layer number one of the temporal pyramid histogram corresponds to Dollár et al.'s work, and may be compared. In the following experiments, we show that action recognition accuracy improves as the layer number increases to an optimal value.

4.1 Facial Expression Dataset

Identical to the experiment setup in (Dollár et al., 2005)'s work, We test the data under two conditions: (1) training on a single subject under one of the two lighting setups and testing on the other subject under the same illumination. (2) training on a single subject under one of the two lighting setups and testing on the other subject under a different lighting setups. The dimension of the spatio-temporal feature point descriptor is 100, and the codebook size is 250.

We vary the layer number L in equation (1) from 1 to 8. Fig. 6 shows the average recognition accuracies on different layers under different experimental setups. Generally, the average recognition accuracy improves when the layer number increases, and drops or remains stable after reaching a peak. The results reveal the fact that sub-clips of the original video do contain useful information for recognition. It can also been concluded that if the temporal layers are too many, recognition accuracy will be not be further improved. This is because too small sub-clips of the original video may not be meaningful for action recognition, and may also introduce noise in the representation.

A comparison between our method and Dollár et al.'s when training and test datasets are under the same

Table 1: Comparison of our result and (Dollár et al., 2005)'s result on two subjects A and B under the same illumination in facial expression dataset.

Method/Accuracy	A/B (same illum.)	B/A (same illum.)
(Dollár et al., 2005)	0.853	0.835
Our approach, layer 1, untuned	0.750	0.708
Our approach, average over 8 layers	0.828	0.758
Our approach, best	0.896	0.813

Table 2: Confusion matrix of our best result, train on subject A/Test on subject B (under the same illumination).

A/B	anger	disgust	fear	joy	sadness	superise
anger	1.0	0.0	0.0	0.0	0.0	0.0
disgust	0.0	1.0	0.0	0.0	0.0	0.0
fear	0.25	0.375	0.375	0.125	0.0	0.0
joy	0.0	0.0	0.0	1.0	0.0	0.0
sadness	0.0	0.0	0.0	0.0	1.0	0.0
suprise	0.0	0.0	0.0	0.0	0.0	1.0

Table 3: Confusion matrix of our best result, train on subject B/Test on subject A (under the same illumination).

B/A	anger	disgust	fear	joy	sadness	superise
anger	1.0	0.0	0.0	0.0	0.0	0.0
disgust	0.0	1.0	0.0	0.0	0.0	0.0
fear	0.0	0.0	1.0	0.0	0.0	0.0
joy	0.125	0.0	0.0	0.875	0.0	0.0
sadness	0.0	0.0	0.0	0.0	1.0	0.0
suprise	0.0	0.0	0.0	0.0	0.0	1.0

illumination settings are shown in Table 1. In Tables 2 and 3 the confusion matrices of our best results are shown, obtained when the training and test datasets are under the same illumination settings. In Tabel 2, fear is misrecognized as anger, disgust and joy. In Table 3, joy is misrecognized as anger. This is because the degree of mouth opening is smilar among these emotions, see Figure 3. From the experiments, we can conclude that the accuracy of our method improves the recognition as the layer number increases. The best accuracy achieved (last row, Table 1) is compareble to Dollár et al.'s result (first row, Table 1).

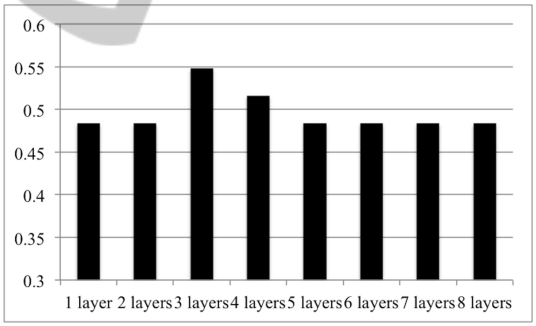

Figure 7: Average recognition accuracy on a subset of UCF-CIL Action dataset.

4.2 UCF-CIL Action Dataset

In our experiment, we include 5 action types of actions, namely golf-swing, one handed tennis backhand stroke, two-handed tennis backhand stroke, tennis forehand stroke and tennis serve from the original dataset. This is because the intuition behind our description approach is to describe not only the codewords but also their temporal sequence, while in repeated actions like fouettes, push up and spin, the temporal sequence of code words is not obvious.

During the experiment, we select one clip from each action to generate the codebook. During testing on the remaining data, we compare every clip to the other clips. The dimension of spatio-temporal feature points is 100. The codebook size is 500.

From Fig. 7, we can conclude that the proposed temporal pyramid histogram description improves the recognition accuracy from 0.48 to 0.55 and the maximum accuracy is achieved for 3 layers. The overall recognition accuracy is not high. This is because two clips of the same action may start from difference poses and end at different ones. This affects the capability of the temporal pyramid histogram descriptor, as it is designed for well-aligned clips ideally. However, we can conclude that temporal pyramid histogram does work on clips that are not well-aligned and also improve accuracy with more layers.

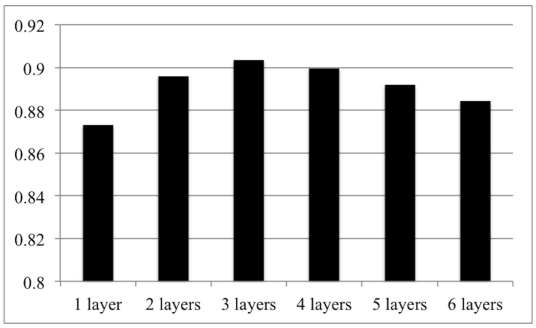

Figure 8: Average recognition accuracy on KTH dataset.

4.3 KTH Human Action Dataset

For the KTH dataset, considering size of the dataset, we obtain the codebook from the videos of three subjects according to the standard codebook generation approach. Then on the remaining 22 subjects, we compute pyramid histograms for 21 subjects, and recognize the last subject by comparing its pyramid histograms to the 21 training pyramid histograms.

The layer number L in equation (1) is varied from 1 to 6. The average recognition accuracy with different layers is in Fig 8. The dimension of spatio-temporal feature points is 100, and the codebook size is 500.

From Fig. 8, we can see that there is a small improvment in recognition accuracy from 0.87 to 0.90 when the layer number changes from 1 to 3. The reason that improvement is not dramatic is that compared to facial expression data, the actions in KTH are repeating actions, for example, walking or running. The proposed pyramid temporal histograms descriptor can not describe the temporal structure of such repeating actions, because each action is performed several times in each video. However, the proposed temporal pyramid histogram description method does enhance the stability of the code words occurrence by calculating the histograms of code words on different layers, as the representative code words of each action occur frequently in all of the subsequences.

(Dollár et al., 2005) obtained over 0.80 on KTH action dataset. According to the evaluation of (Wang et al., 2009) on current spatio-temporal feature point algorithms, the highest recognition accuracy on KTH action dataset based on (Dollár et al., 2005) work is 0.90. Because the parameter settings are not available and the experiment settings are different, the results are not directly comparable. However, the result from the first layer in our approach corresponds to Dollár et al.'s work, and our approach is able to improve the recognition accuracy and match the best.

4.4 Summary

As the baseline parameter settings employed by others for feature extraction and codebook generation are not available, direct comparisons of our method with published results are problematic. However, the results due to (Dollár et al., 2005) correspond to our first layer, therefore our best results, achieved at optimal layers, are equal to or better than the state-of-the-art.

From the results, we can conclude that the temporal pyramid histograms work better when the clips are well-aligned and contain non-repeating actions. However, it can still work when these two conditions are not met.

In the facial expression dataset, because the expression is performed just once in each video and the videos are well-aligned, the proposed description method gets the code words layout correctly. In UCF-CIL action dataset and KTH dataset, the proposed approach also improves the recognition results. This demonstrates the capability of the pyramid histogram descriptor approach in dealing with clips that are not well-aligned or contain repeated actions, though the recognition accuracy cannot be greatly improved. This is because the subsequence segmentation method we employed cannot always pick up meaningful subsequences. In future, we shall explore a more flexible subsequence selection approach to obtain more meaningful action atoms.

5 CONCLUSIONS

In this paper, we introduce a temporal pyramid histograms-based action descriptor approach. Inspired by the spatial pyramid descriptor proposed by (Bosch et al., 2007), we divide the video into subsequences repeatedly, and then concatenate the histograms obtained, from the video parts. To enhance the descriptors ability, we assign different weights to the histogram atoms. The intuition is to encode not only the occurrence of the code words, but also their temporal structure. Although we utilize histogram of code words approach based on the spatio-temporal feature points introduced by (Dollár et al., 2005), it is not hard to conclude that other description approaches, for example, silhouettes, can also be employed in this description framework.

In all our experiments, we obtain improved accuracies as the number of layers is increased to an optimal value, which is the main contribution of the paper. The proposed description approach was tested and found to be most effective in classifying well-aligned clips containing non-repeating actions.

However, this approach does require the videos to be already well segmented and aligned, which means that in each video, there should be only one full performance of an action, i.e. actions of the same type start from the same pose and end at the same pose. In future work, we shall explore automatic subsequence segmentation methods, in order to obtain meaningful subsequences corresponding to action atoms. Therefore, the performance could be improved further on different kinds of datasets. Also, in this work, the weight of each histogram is set by experience, which can be learned automatically in future.

ACKNOWLEDGEMENTS

We would like to thank Dr. Piotr Dollár for generously sharing the feature extraction source code and toolbox.

REFERENCES

Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. (2005). Actions as space-time shapes. In *ICCV 2005*, volume 2, pages 1395–1402 Vol. 2.

Bobick, A. and Davis, J. (2001). The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(3):257–267.

Bosch, A., Zisserman, A., and Munoz, X. (2007). Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 401–408. ACM.

Choi, J., Jeon, W. J., and Lee, S.-C. (2008). Spatio-temporal pyramid matching for sports videos. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, MIR '08, pages 291–297. ACM.

Dollár, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65–72.

Efros, A., Berg, A., Mori, G., and Malik, J. (2003). Recognizing action at a distance. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 726–733 vol.2.

Gilbert, A., Illingworth, J., and Bowden, R. (2009). Fast realistic multi-action recognition using mined dense spatio-temporal features. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 925–931.

Laptev, I. and Lindeberg, T. (2003). Space-time interest points. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 432–439 vol.1.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR2006*, volume 2, pages 2169–2178.

Marszalek, M., Laptev, I., and Schmid, C. (2009). Actions in context. In *CVPR 2009*, pages 2929–2936.

Niebles, J. C., Wang, H., and Fei-fei, L. (2006). Unsupervised learning of human action categories using spatial-temporal words. In *In Proc. BMVC*.

Oikonomopoulos, A., Patras, I., and Pantic, M. (2005). Spatiotemporal salient points for visual recognition of human actions. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 36(3):710–719.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image Vision Comput.*, 28(6):976–990.

Schuldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions: a local svm approach. In *ICPR 2004*, volume 3, pages 32–36 Vol.3.

Scovanner, P., Ali, S., and Shah, M. (2007). A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th International Conference on Multimedia*, pages 357–360. ACM.

Shen, Y. and Foroosh, H. (2008). View-invariant recognition of body pose from space-time templates. In *CVPR 2008*, pages 1–6.

Sun, J., Wu, X., Yan, S., Cheong, L.-F., Chua, T.-S., and Li, J. (2009). Hierarchical spatio-temporal context modeling for action recognition. In *CVPR 2009*, pages 2004–2011.

Wang, H., Ullah, M. M., Klser, A., Laptev, I., and Schmid, C. (2009). Evaluation of local spatio-temporal features for action recognition. In *University of Central Florida, U.S.A*.

Wang, Y., Huang, K., and Tan, T. (2007). Human activity recognition based on r transform. In *CVPR2007*, pages 1–8.

Willems, G., Tuytelaars, T., and Gool, L. (2008). An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proceedings of the 10th European Conference on Computer Vision: Part II*, pages 650–663. Springer-Verlag.