Spatial Multi-Scale Motion History Histograms and its Applications

Asim Jan[1⋆], Zunduo Zhao[2⋆], Tong Chen[2], Hongying Meng[1,2], and Tao Lei[3]

[1] Brunel University London, UK,
[2] Southwest University, Chongqing, China,
[3] Shaanxi University of Science and Technology, Xi'an, China

Abstract. Precisely describing the action inside of a video is a challenging task because the content of the video includes various objects, with different local motion information at different speed in the video frames. In this paper, a new video feature is proposed based on the spatial information of the objects in a frame, along with the motion information between one against multiple consecutive frames. Motion information between pixels at the same position in the whole video are all combined for a new Spatial Multi-Scale Motion History Histogram (SMMHH) dynamic descriptor. The detailed algorithm of the SMMHH was given and it is tested in both human action recognition and touch gesture recognition applications based on the public video datasets. Experimental results demonstrate its excellent performance compared to other traditional methods.

Keywords: Motion, Video Dynamic Feature, Human Action Recognition, Hand Gesture Recognition

1 Introduction

Dynamic video descriptor provides an efficient description of the occurring dynamic actions. It is helpful for automatic systems in understanding and analyzing video content and can be further applied in the applications of human-computer interaction, robotics and automatic multi-media content analysis for big complex data. In order to capture the dynamic information in the video, video frames are compared, and the motion information are extracted between frames and a statistical feature is extracted. The two typical features are Motion History Histogram (MHH) [1] and Local Binary Patterns on Three Orthogonal Planes (LBPTOP) [2] that are applied in many applications with great success. In this paper, MHH feature is further studied. Especially, the spatial and time do-mains are further investigated to extract motion occurring across different frames. These extensions keep the basic idea of MHH and take the advantages of introducing different motion dynamics.

⋆ Equal contribution

2 Related Works

For appearance-based methods, there has been a continuous effort in the computer vision society for extracting dynamic descriptor such as Motion Energy Images and Motion History Image [3]. The feature itself lacks the depth of information as it is solely based on the energy differences between two frames at a time. A more effective approach is to summarize the energy across a whole sequence as a single vector. This issue was addressed by Weinland et al. [4] as they extend the 2D motion templates to create Motion History Volumes. They claim that a 3D representation of motion templates is a more robust and natural way to fuse information from multiple images, producing view-invariant features.

There are also techniques that look for patterns and edges within the temporal space. LBP-TOP is a method that works well, which is based on the popular spatial method Local Binary Patters (LBP) [5]. Zhao and Pietikainen [2] introduced their dynamic LBP operator for texture recognition and for facial expressions and human action recognition [6]. A variant on LBP-TOP has been developed by Almaev and Valstar [7] called Local Gabor Binary Patterns-Three Orthogonal Planes that is an extension of Local Gabor Binary Patterns produced by Senechal et al [8]. Local Phase Quantization - Three Orthogonal Planes (LPQ-TOP) is another descriptor for temporal data, similar to LBP-TOP, by capturing LPQ features across the XY, XT and YT dimensions. This is developed by Jiang et al. [9] also for detecting facial action units. Histogram of Oriented Gradients 3D (HOG3D) [10] created by Klaser, Marszaek and Schmid is a descriptor based on histograms of oriented spatio-temporal gradients, based on Histogram of Oriented Gradients (HOG) [11].

These techniques do not look for ways to improve how the spatial and time information can be handled, in a way that would produce higher quality features based on temporal movement. This paper expands on the MHH technique factoring in ideas to make better use of the spatial and time information.

3 Motion History Histogram and Its Extensions

3.1 Basics of Motion History Histogram

MHH [12] captures the motion from a video in its gray-scale form, this is done by detecting how much each pixel moves across the frames with further details available in [1] [13] and its algorithm is shown below. Let $\{f(u, v, k),\ u = 1, \cdots, U;\ v = 1, \cdots, V;\ k = 1, \cdots, K\}$ be a video clip where k is the frame number and u, v are the row and column of the pixels. We define $\{D(u, v, k),\ k = 1, \cdots, K\}$ as the binary sequence on pixel (u, v) that is computed by firstly calculating absolute difference differences between frame $k + 1$ and frame k and then comparing it to a Threshold T. $I(u, v)$ is a frame index that stands for the number of the starting frame of a new pattern on pixel (u, v). At the beginning, $I(u, v) = 1$ for all (u, v). That means a new pattern starts from frame 1 for every pixel. $I(u, v)$

Algorithm 1 Algorithm (MHH)

Input: Video clip $f(u, v, k), u = 1, \cdots, U, v = 1, \cdots, V, framek = 1, \cdots, K$
Initialisation:
Possible patterns: $i = 1, , M$,
MHH$(1 : U, 1 : V, 1 : M) = 0$,
Pattern starting index $I(1 : U, 1 : V) = 1$
1: **for** $k = 2$ to K (For 1) **do**
Compute: $D(:, :, k)$ (0 or 1 based on frame difference)
2: **for** u=1 to U (For 2) **do**
3: **for** v=1 to V (For 3) **do**
4: **if** $D(u, v, k) = 0$ (If 1) **then**
5: **if** $D(u, v, I(u, v)), , D(u, v, k)$ is pattern i (If 2) **then**
Update: MHH(u, v, i) =MHH$(u, v, i) + 1$
6: **end if**
Update: $I(u, v) = k$
7: **end if**
8: **end for**
9: **end for**
10: **end for**
Output MHH$(1 : U, 1 : V, 1 : M)$

will be updated to $I(u, v) = k$ while $D(u, v, I(u, v)), \cdots, D(u, v, k)$ builds one of the patterns $i(1 \leq i \leq M)$ and, in this case, MHH(u, v, i) increases by 1.

3.2 Multi-Scale Motion History Histogram

Multi-scale Motion History Histogram (MMHH) is an extension to MHH which can provide distinction between faster and slower motion movement. The approach here is to compute the motion across different subsets of frames, which can be achieved by frame skipping at different levels during the motion detection process. The algorithm can be implemented as follows:

$$G(u, v, k, s) = |f(u, v, k + s) - f(u, v, k)| \tag{1}$$

$$D(u, v, k, s) = \begin{cases} 1, & \text{if } G(u, v, k, s) \geq T \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

$$MMHH(u, v, m, s) = \begin{cases} MMHH(u, v, m, s) + 1, & \text{if } P_m \text{ is found} \\ MMHH(u, v, m, s), & \text{otherwise} \end{cases} \tag{3}$$

Using a similar approach of MHH, $G(u, v, k, s)$ represents the absolute difference between the frames $f(u, v, k + s)$ and $f(u, v, k)$. $D(u, v, k, s)$ represents the binary sequence on pixels (u, v), and MMHH is the resulting histogram of the patterns $P_{1:M}$. Multi-Scale now uses the threshold to compare the differences between frame $k + s$ and frame k, where $s = 1, ..., S$ represents the Scale level. The full algorithm for this can be seen in Algorithm 2. Varying the scale value S effectively results in capturing motion at different speeds, which can be seen in Fig. 1.

Algorithm 2 Algorithm (MMHH)

Input: Video clip $f(u, v, k), u = 1, \cdots, U, v = 1, \cdots, V, framek = 1, \cdots, K$
Initialisation:
Possible patterns: $i = 1, , M$,
Scales: $s = 1, , S$,
MMHH$(1 : U, 1 : V, 1 : M, 1 : S) = 0$,
Pattern starting index $I(1 : U, 1 : V) = 1$
for s=1 to S (For S) **do**
  for k=2 to K (For K) **do**
Compute: $D(:, :, k, s)$ (0 or 1 based on frame difference)
    for u=1 to U (For U) **do**
      for v=1 to V (For V) **do**
        if $D(u, v, k, s) = 0$ (If 1) **then**
          if $D(u, v, I(u, v), s), , D(u, v, k, s)$ is pattern i (If 2) **then**
            Update: MMHH(u, v, i, s) =MMHH$(u, v, i, s) + 1$
          end if
Update: $I(u, v) = k$
        end if
      end for
    end for
  end for
end for
Output MMHH$(1 : U, 1 : V, 1 : M, 1 : S)$

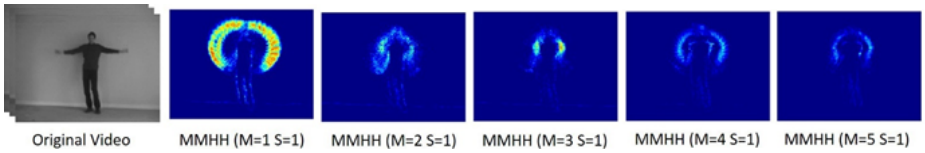

Fig. 1. Example based on a sample from the KTH dataset of a person waving their hands of MMHH where $M = 1 : 5$ and showing only at multi-scale $S = 1$.

3.3 Spatial Motion History Histogram

Spatial pooling is another concept introduced to MHH to provide a different view of visual data, taken from the idea of Convolutional Neural Networks (CNNs), to produce Spatial Motion History Histogram (SMHH).
There are 3 different parameters available to configure for the spatial pooling stage. α determines the reduction factor of each frame from a video. $L(x)$ is the binary operator applied during the spatial pooling, which lets the user select between calculating the average ($x = 0$) or max ($x = 1$) value of the pixels. Finally, N determines the number of times a video will have spatial pooling applied.

$$PV(1 : U/\alpha, 1 : V/\alpha, 1 : K) = Poolf(u, v, k), L(x) \tag{4}$$

$$Pool = \begin{cases} \text{AVERAGE}(f(u : u + \alpha, v : v + \alpha, k)), & \text{if } \{\text{L(x)==0}\} \\ \text{MAX}(f(u : u + \alpha, v : v + \alpha, k), & \text{otherwise} \end{cases} \tag{5}$$

Algorithm 3 Pooling for SMMHH

Input: Video clip $f(u,v,k), u = 1,\cdots,U, v = 1,\cdots,V, framek = 1,\cdots,K$ Parameters x, α
Initialisation:
PoolType $= x$, PoolSize $= \alpha$,
$PV(1:U/\alpha, 1:V/\alpha, 1:K = 0$,
for $k = 1,\cdots,K$ (For K) **do**
  for $u = 1,\cdots,U/\alpha$ (For U) **do**
    for $v = 1,\cdots,V/\alpha$ (For V) **do**
      if $x = 0$ **then**
Cal: P $= AVERAGE(f((u-1)\times\alpha+1:u\times\alpha,(v-1)\times\alpha+1:v\times\alpha,k))$
Update: $PV(u,v,k) = P$
      else if $x = 1$ **then**
Cal: P $= MAX(f((u-1)\times\alpha+1:u\times\alpha,(v-1)\times\alpha+1:v\times\alpha,k))$
Update: $PV(u,v,k) = P$
      end if
    end for
  end for
end for
Output $PV(1:U,1:V,1:K)$

Equation 4 produces $\{PV(u,v,k), u = 1,\cdots,U/\alpha; v = 1,\cdots,V/\alpha, k = 1,\cdots,K\}$, which is the new pooled video. Pool is the operator in equation 5 that calculates PV based on Average or Max pooling. Each video can be pooled N times, based on the frame dimensions and the reduction factor applied every time the data is spatially pooled. MHH is then applied on each of the resulting transformed temporal data. As the resulting histograms from each pooled data are of different dimensions, a final feature vector can be made by reshaping then concatenating the pattern frames M for each MHH feature, or by extracting local features of every pattern frame M .

3.4 Spatial Multi-Scale Motion History Histogram

With the two extensions MMHH and SMHH, both efforts can be combined to make a large feature called Spatial Multi-Scale Motion History Histogram (SMMHH). Initially the spatial pooling is applied on the original temporal data $f(u,v,k)$ N times to produce $N+1$ videos. Then for each transformed data, MMHH is captured and concatenated to provide the SMMHH feature. The full algorithm for Pooling can be seen in Algorithm 3 and the overall process can be seen in Fig. 3.

4 Experimental Result

4.1 Human Action Recognition

This experiment is an attempt to classify 6 Human action gestures using the KTH dataset by Schuldt et al. [14].

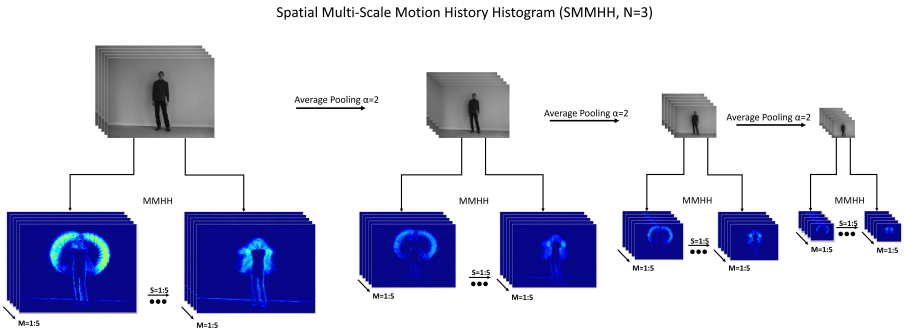

Fig. 2. Overview of SMMHH being applied on a person waving their hands. Average pooling is applied $N = 3$ times on the temporal data, $\alpha = 2$, with Multi-scale of $S = 1 : 5$ with all patterns from $M = 1 : 5$.

We then split the subjects into 16 for training and 9 for testing as done in [14] and [13]. The proposed methods are applied on the first 300 frames of the 599 video clips. Three machine learning methods have been chosen to compare the performance of the features, which include SVM, RF and KNN.

Table 1. KTH Dataset: Human Action Recognition Experiment Results

Feature Set	SVM	RF	KNN
LBP-TOP[6]	86.25%	-	-
Extended LBP-TOP[6]	88.19%	-	-
Ext Grad LBP-TOP[6]	91.25%	-	-
HOF-HOG[15]	89.88%	-	-
HOG3D[10]	91.40%	-	-
MHH + EOH	85.64%	86.11%	80.09%
SMHH(M) + EOH	83.80%	79.63%	70.89%
SMHH(A) + EOH	86.57%	87.11%	84.72%
MMHH + EOH	88.42%	81.94%	81.48%
SMMHH(M) + EOH	87.96%	82.87%	77.78%
SMMHH(A) + EOH	**92.12%**	**87.44%**	**88.89%**

Table 1 shows the performance on all the methods using the same setup. When comparing to other features, SMMHH(A) EOH using SVM outperforms the Extended Gradient LBP-TOP and HOG3D features by 0.87% and 0.72% respectively.

Table 2. Cost Dataset: Accuracies for different feature sets

Data Set	Feature Set	Experiment Setup	RF	SVM	KNN
Training	MHH	10 Fold CV	54.30%	46.23%	41.90%
	MMHH	10 Fold CV	59.80%	54.20%	47.39%
	SMMHH(M)	10 Fold CV	60.35%	56.40%	48.40%
	SMMHH(A)	10 Fold CV	**60.80%**	**56.70%**	**52.40%**
	LBPTOP	10 Fold CV	55.63%	50.80%	40.80%
Testing	SMMHH(A)	10 Fold CV	**61.25%**	**56.52%**	**51.78%**
	LBPTOP	10 Fold CV	55.99%	52.91%	38.05%
Testing	SMMHH(A)	Challenge	**54.60%**	**53.37%**	**38.18%**
	LBPTOP	Challenge	50.86%	47.59%	25.49%

4.2 Social Touch Gesture Recognition

This experiment is based on the Touch Challenge 2015 at the 17th ACM International Conference on Multi-modal Interaction, for recognizing touch gestures amongst humans. The Corpus of Social Touch (CoST) dataset is based on touch gestures with human interaction [16]. It contains 14 types touch gestures. The training is assessed using subjects randomized, with 10-Fold cross-validation to get a reliable classification rate. The machine learning models used in this experiment are consisted of RF, SVM and KNN. We applied Spatial Pooling with $N = 3$, and a reduction factor $\alpha = 2$, which produce a total of 4 videos with frame sizes: 8x8, 4x4, 2x2 and 1x1 resolutions. On each of the videos we then extract the MMHH feature with $S = 5\&M = 5$, with an appropriate threshold. All of the features produced are then reshaped and concatenated to a feature vector with the dimension of $\sum_{i=1}^{4}(S \times M \times P_i) = 2125$. where i refers to each of the Spatially Pooled videos, Pi is the total number of pixels in a frame for each video i, $S = 5$ is the size of the Multi-Scale dimension, which has been set from 2:6 and $M = 5$ is the MHH pattern sequence size [12]. Table 2 contains the results for the CoST dataset. It can be seen that SMMHH(A) has performed significantly better (6.5% across 14 classes) than the original MHH feature. MMHH and SMMHH(M) have also shown to be more successful than MHH and LBP-TOP showing that the proposed variations do provide richer information for distinguishing across many classes.

5 Conclusion

In this paper, a new SMMHH feature is proposed and evaluated in two public datasets for different applications. Experimental results show that SMMHH feature achieved better results than original features and popular LBPTOP feature and other variants. The extensions MMHH, SMHH and SMMHH provide a new scale of information by looking at motion differently, using the time and spatial domains. It can be a general video dynamic feature for many applications.

References

1. Hongying Meng, Nick Pears, Michael Freeman, and Chris Bailey. *Motion history histograms for human action recognition*, pages 139–162. Springer, 2009.
2. Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis Machine Intelligence*, (6):915–928, 2007.
3. James W Davis and Aaron F Bobick. The representation and recognition of action using temporal templates. In *IEEE conference on computer vision and pattern recognition*, pages 928–934.
4. Daniel Weinland, Remi Ronfard, and Edmond Boyer. Free viewpoint action recognition using motion history volumes. *Computer vision and image understanding*, 104(2-3):249–257, 2006.
5. Timo Ojala, Matti Pietikinen, and Topi Menp. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis Machine Intelligence*, (7):971–987, 2002.
6. Riccardo Mattivi and Ling Shao. Human action recognition using lbp-top as sparse spatio-temporal feature descriptor. In *International Conference on Computer Analysis of Images and Patterns*, pages 740–747. Springer.
7. Timur R Almaev and Michel F Valstar. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 356–361. IEEE.
8. Thibaud Senechal, Vincent Rapp, Hanan Salam, Renaud Seguier, Kevin Bailly, and Lionel Prevost. Facial action recognition combining heterogeneous features via multikernel learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(4):993–1005, 2012.
9. Bihan Jiang, Michel F Valstar, and Maja Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Face and Gesture 2011*, pages 314–321. IEEE.
10. Alexander Klaser, Marcin Marszaek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*, pages 275: 1–10. British Machine Vision Association.
11. Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *international Conference on computer vision Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE Computer Society.
12. Hongying Meng, Nick Pears, and Chris Bailey. A human action recognition system for embedded computer vision application. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6. IEEE.
13. Hongying Meng and Nick Pears. Descriptive temporal template features for visual motion recognition. *Pattern Recognition Letters*, 30(12):1049–1058, 2009.
14. Ivan Laptev and Barbara Caputo. Recognizing human actions: a local svm approach. In *null*, pages 32–36. IEEE.
15. Ivan Laptev, Marcin Marszaek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *CVPR 2008-IEEE Conference on Computer Vision Pattern Recognition*, pages 1–8. IEEE Computer Society.
16. Merel M Jung, Ronald Poppe, Mannes Poel, and Dirk KJ Heylen. Touching the void–introducing cost: corpus of social touch. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 120–127. ACM.