Temporal Key Poses for Human Action Recognition

Abdalrahman Eweiwi[1] Shahzad Cheema[1] Christian Thurau[1, 2] Christian Bauckhage[1, 2]
[1]Bonn-Aachen International Center for IT, University of Bonn [2] Fraunhofer IAIS, Germany
{eweiwi, cheema}@bit.uni-bonn.de {christian.thurau, christian.bauckhage}@iais.fraunhofer.de

Abstract

In this paper, we present a simple yet effective approach to recognizing human activities from video sequences. Our approach integrates the advantages of human action recognition in static images using action key poses and motion based approaches using the variants of Motion History Images (MHI) and Motion Energy Images(MEI). We combine both methodologies to extract a new representation of temporal key poses. In an evaluation of this on well established benchmark data we achieve high recognition rates. For the task of action recognition using the MuHAVi data set, we achieve an accuracy of 98.5% in a leave-one-out cross validation procedure. For single-view action recognition using the popular Weizmann data sets, we achieve an accuracy of 100%. In more difficult evaluation setups where the number of training samples for certain individuals or views are restricted, the proposed method exceeds recently published results of other approaches. Moreover, the introduced approach is computationally efficient, robust with respect to parameter selection, and straight forward to implement as it builds on well established and understood concepts.

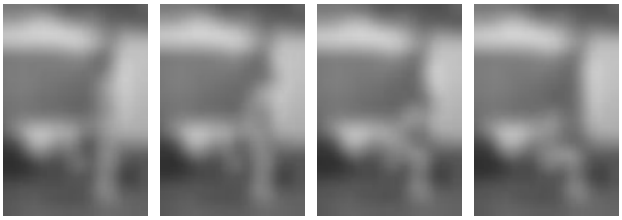
Figure 1. Exemplary silhouettes from the MuHAVi data set (first row) and the Weizmann data set (second row).

1. Introduction

Automatic recognition of human actions is an important and active field in computer vision; it aims at automatically identifying human actions in images or videos. That is, it attempts identifying whether a person is walking, jumping, or performing other types of actions. There are many tangible applications of activity recognition, ranging from domains such as surveillance, video annotation, or gesture recognition to sports video interpretation.

Among the many different approaches to action recognition, two general trends are to consider static, pose-based cues or dynamic, motion-based information. For the pose-based methods, often a sequence of certain expressive poses, so called *key poses* or *pose primitives*, is extracted at a frame level. These methods are limited by the discriminative power of the extracted key poses, specifically, for classes where the inter-class variations is smaller than the intra-class variation, such as the identification of, *e.g.* walking or running. Moreover, these methods can be negatively affected by sudden shape deformations due to noise or occlusions (Figure 1) or the absence of a clearly structured human pose (Figure 2). For the temporal methods, many authors consider simple motion features which are extracted from a complete sequence or subsequences thereof. Interestingly, a direct combination of static pose features and motion feature on frame level has been rarely applied so far [12, 8].

[12] presents an implementation of a biologically inspired system that combines shape- as well as motion-based representations. Shape represent silhouettes and motion is characterized by the optic flow between successive frames. They empirically prove that the combination of both features outperforms corresponding single cue activity recognition. Following this work, [8] derives a pose descriptor for each frame of a video. It captures both motion (from optical



Frame 5 Frame 15 Frame 25 frame 35

Figure 2. Selected frames showing a person in action. Despite the absence of a clearly visible human pose, people still can recognize the activity as somebody sitting down [2].

flow) and pose information (from image gradients). Using clustering, a codebook of visual poses is created and used for later recognition steps. Tested against various benchmark data, this setup achieves competitive results.

Given the known merits of temporal templates of Motion History Images (MHI) and Motion Energy Images (MEI) [2, 5] and the benefits of combining motion and pose information [12], we here propose to apply MHIs, which represent significant motion cues, in combination with representations of static body poses to discriminate different actions. This representation thus integrates well established methods in a new way that alleviates their individual shortcomings.

For the results discussed in this paper, we assume a working background subtraction to be given such that silhouette images are available for further processing. We first localize the human silhouette within a window centered around the center of mass of the shape pixels. Instead of considering one temporal template (MHI) for the whole image sequence, we use a temporal pose template for each body pose at each frame of the action sequence. Next, we perform clustering using k-means which results with a set of temporal key poses. For recognition, a nearest neighbor procedure is applied to determine a class label for each temporal pose template of the query video. Finally, a majority voting scheme is used to determine a class label for the whole image sequence.

In the reminder of this paper, we first discuss related work and then review the idea of temporal templates of MHIs and MEIs. Section 4 presents the data sets used in the evaluation procedure. In section 5, we present our experimental evaluation. Finally we present our conclusions in section 6.

2. Related Work

Automatic event detection from video has been studied extensively in recent years [10, 19]. Roughly, one can categorize proposed algorithms into three different strategies based on the nature of the features used for classification: *dynamic features* (i.e motion cues), *static features* (i.e. shape cues), or implicit or explicit combinations of both. Most recent approaches rely on identifying dynamic features using motion cues. Early attempts to solve the problem of human action recognition used tracked body parts as input features [11, 20]. However, practically feasible representations were impeded by the challenging problem of tracking human body parts, especially in monocular videos.

Interest points methods which are popular in object detection have also found their way to action recognition. [9] presented a hierarchical model which applied a collection of spatial and spatiotemporal features extracted from interest points. [6] extended spatial interest points to 3D space-time interest points in order to extract structures that have significant local variation in both space and time. The resulting representations are then applied to human action recognition using Support Vector Machines.

Methods which depend only on static features have evolved recently, most likely due to generalization capability of this idea which applied to both still images and videos. [15] represents activities as a temporal sequence of key poses. These key poses result from clustering the training data and considering the resulting cluster centroids as discriminative elements. [4] models human poses in videos using the *Histogram of Oriented Rectangles* (HOR) descriptor which describes a pose as a set of oriented rectangles that cover the human body. These methods are yet limited by the discriminative power of those static features, specially, for actions which share strong relations between their poses.

Methods which implicitly combine motion and appearance information date back to the analysis of spatiotemporal templates [2]. Typically, MEI and MHI are used as temporal templates to recognize human actions. For recognition, [2] uses seven Hu moments. [18] proposed a 3D extension of temporal templates. The idea is to use multiple cameras in order to build motion history volumes and to classify actions using Fourier analysis in cylindrical coordinates. [3] proposes a temporal-state shape context (TSSC) method that organizes silhouettes of objects in a video into three temporal states. The objective is then to capture local characteristics of the space-time shape induced by consecutive changes of silhouettes. These approaches rely majorly on modeling the action sequence as a whole instead on short action snippets [12], which may limit any temporal segmentation procedure if the person's behavior changes from one action to another along the sequence.

[8] proposes a pose descriptor capable of capturing both motion (from optical flow) and pose information (from image gradients). Afterwards a clustering scheme is utilized to extract a codebook of visual poses for recognition. The success of combined features was made popular by [12]. The approach combines static form and dynamic motion

features on a frame level. For form, [12] extracts local edges and for motion, the optic flow computed between consecutive frames $t-1$ and t is regarded. Both features are separately compared to previously learned templates. This combination of both features led to an important observation, which identifies the required number of frames required to recognize an action.

In contrast to prior work with temporal templates, [5] models the image sequence of an action as a sequence of temporal templates instead of one for the whole action. Afterwards, a *local binary pattern* (LBP) texture descriptor is used to capture the essential information of the human movements. This approach operates on both image data and silhouette image sequences, and presents competitive results on benchmark datasets. Recently, [16] proposed a simple yet effective approach that aims at encoding human actions using the quantized vocabulary of averaged silhouettes. These silhouettes are derived from space-time overlapping window shapes and implicitly capture local temporal motion as well as global body shape. However, this approach may have a limited applicability toward discriminating actions which involve similar poses in reverse order (e.g. stand up and sit down) since it does not give enough evidence of the chronicle order of the human poses within the selected window.

3. Action Modeling

In work reported here, we model human actions as sequences of temporal templates. We represent each body pose within a sequence by its actual time context. This is performed by transforming each pose to a temporal representation using Motion History Image (MHI) as proposed by [2]. Intuitively, this representation is able to characterize motion and can also reference the underlying static pose of the body [2, 5]. Figure 3 illustrates examples of corresponding temporal pose representations of selected frames of a walking action. The major benefit of transforming static human poses into temporal description is to alleviate the intra-class variation within the same action and to enlarge the inter-class variation between different actions by relating the pose to its chronicle order. In the following, we describe the construction used to extract the temporal pose templates as well as the classification approach used for action recognition.

3.1. MEI and MHI Templates

[2] introduced Motion Energy Images (MEI) and Motion History Images (MHI). These representations decompose motion recognition by describing where object motion is observable (MEI) and how an object is moving (MHI).

An MEI E is a binary construct which identifies where motion occurred in an image sequence (computed from the first to the final frame). In particular, it describes the spatial

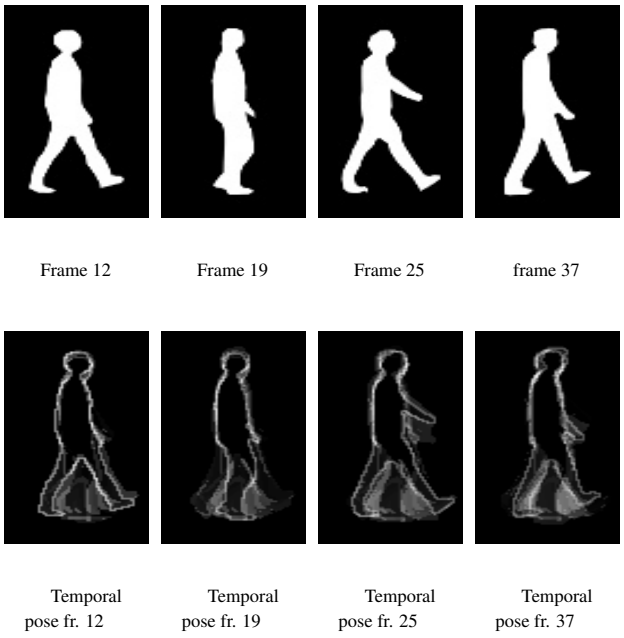


Figure 3. Frames from the walking action in the MuHavi data (first row) and their temporal pose version using MHI (second row).

distribution of motion and is defined as:

$$E_\tau(x,y,t) = \bigcup_{i=0}^{\tau-1} D(x,y,t-i) \tag{1}$$

where the $D(x,y,t)$ are successive differences between video frames, i.e

$$D(x,y,t) = \mid I(x,y,t) - I(x,y,t-1) \mid \tag{2}$$

where, $I(x,y,t)$ refers to a (binary) silhouette image with coordinate (x,y) at time t.

On the other hand, an MHI H captures the spatial and temporal information of motion in images, encoding how recent changes have occurred. If motion is present at time t, i.e $D(x,y,t) = 1$, then each pixel motion of the MHI is a function of the history of motion at that point, occurring within a fixed duration τ:

$$H_t(x,y) = \begin{cases} \tau & \text{if } D(x,y,t) = 1 \\ \max(0, H_{t-1}(x,y) - \delta) & \text{otherwise} \end{cases} \tag{3}$$

Here (x,y) and t denote position and time, while D signals an object's presence in the current video image. The duration τ denotes the temporal extent of a motion and δ represents a decay parameter. For simplicity, we set $\delta = 1$. In the resulting grey-value image, more recently moving pixels appear brighter than older ones [2].

Grey-value MHIs are able to distinguish between mirror symmetric actions (*e.g.* "walk left" vs. "walk right"). In

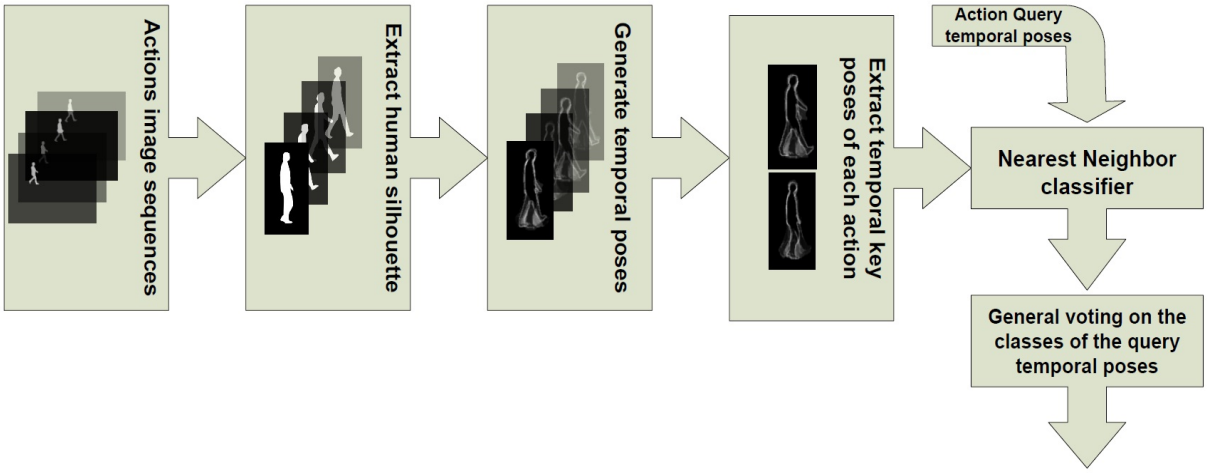


Figure 4. General overview of our approach. First human silhouettes are extracted, resized, and centered around their center of mass. Afterwards, temporal pose templates are extracted using MHIs and clustered separately for each action. Cluster centroids (temporal key poses) are then used to classify temporal poses of novel image sequences using a nearest neighbor scheme. A majority voting over the resulting per frame labels of query sequence decides the final class label of the sequence.

addition, MHI templates are insensitive to silhouette noise such as holes, shadows, or missing parts. They can be computer from cheap cameras and on less powerful CPUs and work well where structure cannot be easily detected [1].

MHIs are known to be sensitive against changes in scale, location, and view point. We eliminate scale and location dependencies by centering each MHI with respect to the center of mass of the human silhouette [14]. Another known shortcoming of drawback of MHIs occurs when they are used to represent an action as a whole. Therefore, setting the duration parameter τ is critical, since the duration of different actions varies (*e.g*. kicking and running), as do different instances of the same action. In our implementation, we address this by using a temporal template for each frame instead of using one for a whole video sequence.

3.2. Classification

Our classification approach applies a simple and efficient yet effective nearest-neighbor classifier. In order to train it from raw image sequences, the extracted human silhouettes are resized to a common size and centered around their center of mass. Afterwards we extract for each pose in a activity video its MHI template. This results with a large number of templates, most of which contain redundant and insubstantial information. To focus on the most discriminative templates of each action, we resort to k-means clustering on the training data of each action separately. We refer to the resulting cluster centroids as *temporal key poses*. In the recognition phase, given novel sequences, we first generate their temporal poses templates and assign them to the class of the nearest template key pose previously learned. The final decision of the action class results from a majority vote over all pose templates classes of the query video.

4. Data Sets

Our evaluation experiments were performed on two benchmark data sets, namely, the Weizmann data set, and the MuHAVi data set. For the Weizmann data set, we consider the silhouette images of the aligned version for the human action images. These silhouettes contain ”leaks” and intrusions” due to shadows and imperfect subtraction with the background. The silhouette masks provided consist of 93 samples of 9 different people, each performing 10 natural actions such as ”run”, ”walk”, ”skip”, ”wave”, ”jump”, ”bend”, etc. All actions were recorded from the same viewpoint in a controlled environment.

The Multiple-view Human Action Video data set (MuHAVi) data set has recently raised the bar for the challenge of human action recognition. It provides multi-view data of actions of different actors with CCTV-like views (at an angle and some distance from the observed person). Action videos in this data set were taken from multiple view points performed by multiple actors. The data consists of 136 samples of 14 primitive actions performed by 2 actors observed from 2 different views. The actions in the data set can be reorganized into 8 classes where similar actions now constitute a single class. In [13], the authors have set the baseline accuracies for three challenging evaluations. The first is to perform a leave one out evaluation on the whole data set; the second considers identical camera views for training and testing, yet testing happens with videos of actors not contained in the training data. The final evaluation

Table 1. Accuracy(%) Leave one out, **8 actions**

	Accuracies		
Action class	[13]	Our	[7]
Collapse	16/16	16/16	16/16
Run	15/16	16/16	15/16
Standup	12/12	12/12	12/12
TurnBack	11/12	12/12	12/12
Walk	16/16	15/16	16/16
Guard	32/32	32/32	30/32
Kick Right	15/16	15/16	16/16
Punch Right	16/16	16/16	16/16
Average Accuracies	97.8 %	**98.5%**	97.8%

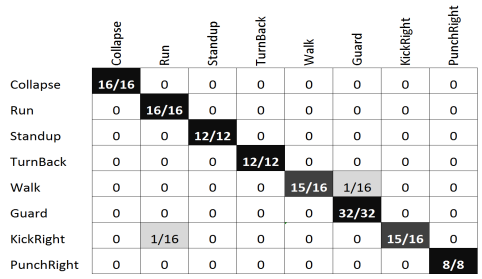


Figure 5. Confusion matrix for leave-one-out cross evaluation on **8 actions** from the MuHavi data set (average accuracy 98.5% with 2 samples out of 136 misclassified).

setup tests identical training and test actors against videos recorded from novel camera views.

5. Experiments

We perform an extensive evaluation of our proposed approach on each of the above mentioned data sets. In case of the MuHAVi data set, we perform the suggested 3 tests: leave-one-out evaluation on the whole data set, novel actor evaluation, and novel camera viewpoint evaluation. We also perform further evaluations on the Weizmann data set, in order to insure the validity of our algorithm, and compare it to known state-of-the-art results.

5.1. MuHAVi Leave-one-out Cross Evaluation

For the first setup, we consider the MuHAVi data set composed of **8 different actions**. We perform a leave-one-out cross-validation. We achieve an accuracy rate of 98.5%, i.e. only 2 out of 136 samples were misclassified. Table 1 illustrates the individual accuracy rates obtained for the 8 actions. Note that, in the corresponding confusion matrix in Figure 5, confusion occurs between the "guard-to-kick" and "guard-to-punch" actions which we attribute to the high visual similarity of these actions.

Table 2. Accuracy(%) Leave one out, **14 actions**

	Accuracies	
Action class	[13]	Our
Collapse Left	4/8	6/8
Collapse Right	5/8	7/8
Run Left To Right	7/8	8/8
Run Right To Left	7/8	8/8
Stand up Right	8/8	8/8
Turn Back Right	7/8	8/8
Walk Left To Right	8/8	8/8
Walk Right To Left	8/8	8/8
Guard To Kick	13/16	12/16
Guard To Punch	10/16	14/16
Kick Right	15/16	16/16
Punch Right	16/16	16/16
Turn Back Left	4/4	2/4
Stand up Left	0/4	4/4
Average Accuracies	82.4%	**91.9%**

Next we perform similar evaluation on the MuHAVi data using a leave-one-out cross-validation setup, but this time consider **14 different primitive actions**. Table 2 shows the resulting accuracies. Similar to the setup with 8 action classes, we achieve an accuracy rate of 91.9%, i.e 11 samples out of 136 are misclassified. Figure 6 provides the corresponding confusion matrix.

5.2. Identical Cameras, Novel Actors

In this setup, we split the data set into 2 parts each containing 68 samples. Each part refers to actions performed by the same actor. For classification, we train with one actor actions and test it against another. In an 8 action setup, we achieved an accuracy rate of 85.3%. Table 3 illustrates the accuracy rates achieved for each class.

Apparently, our approach outperforms the baseline method with 10 misclassified samples compared to 16 in [13]. See Figure 7 for the confusion matrix.

Again, we perform similar actor-to-actor evaluation for the data set consisting of 14 primitives actions. In this test, too, accuracies outperform the baseline approach [13] reaching a rate of 77.9% The number of misclassified samples was found to be 15. Table 4 illustrates summarizes the results; Figure 8 provides the confusion matrix.

5.3. Identical Actors, Novel Camera

We perform this test to verify the performance of our algorithm on various views of human actions. For this purpose, we group the data into two parts, each containing 68 samples of the actions performed by one actor from one view. We train our classifier on one view and test against

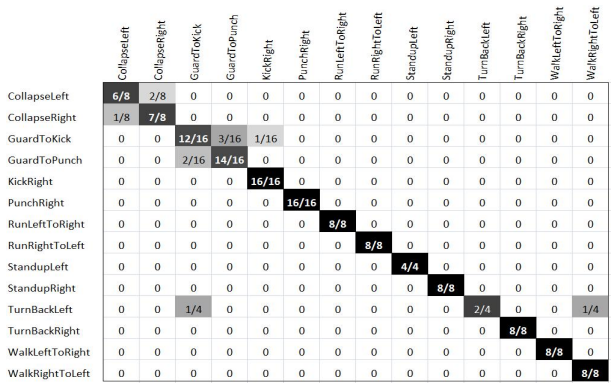


Figure 6. Confusion matrix for leave-one-out cross evaluation on **14 actions** from the MuHavi data set (average accuracy 91.9% with 11 samples out of 136 misclassified).

Table 3. Accuracy(%) Novel actors, **8 actions**

	Accuracies	
Action class	[13]	Our
Collapse	6/8	3/8
Run	7/8	8/8
Standup	5/6	6/6
TurnBack	3/6	5/6
Walk	8/8	7/8
Guard	8/16	13/16
Kick Right	7/8	8/8
Punch Right	8/8	8/8
Average Accuracies	76.4 %	**85.3%**

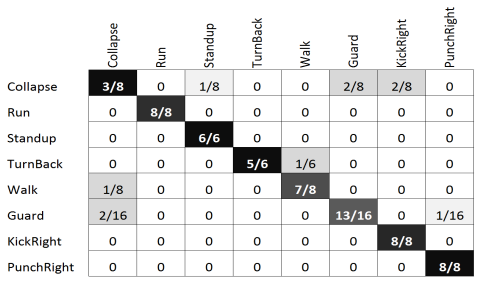


Figure 7. Confusion matrix for novel actor evaluation, **8 actions**

another. For the 8 actions setup, we achieved an accuracy rate of 55.8%, i.e 30 samples out of 68 are being misclassified (Table 5). This result is anticipated because of the viewpoint sensitivity of the MHI as pointed out by [2]; Figure 9 provides the confusion matrix.

For the 14 actions setup, we achieved a lower accuracy rate of 38.2% with 42 misclassified samples out of 68 (Table 6). This is a reasonable result when considering 14 actions, since we can attribute it to the MHI representation eliminating mirror symmetries. For instance, actions such as "walk left" recorded by a camera 3 would appear as "walk right" from the point of view of another camera. Artifacts like these require special attention in multi-view

Table 4. Accuracy(%) Novel actors, **14 actions**

	Accuracies	
Action class	[13]	Our
Collapse Left	3/4	2/4
Collapse Right	2/4	1/4
Run Left To Right	3/4	4/4
Run Right To Left	4/4	4/4
Stand up Right	4/4	4/4
Turn Back Right	2/4	4/4
Walk Left To Right	4/4	4/4
Walk Right To Left	4/4	3/4
Guard To Kick	0/8	2/8
Guard To Punch	0/8	7/8
Kick Right	0/8	8/8
Punch Right	7/8	7/8
Turn Back Left	1/2	1/2
Stand up Left	0/2	2/2
Average Accuracies	61.8%	**77.9%**

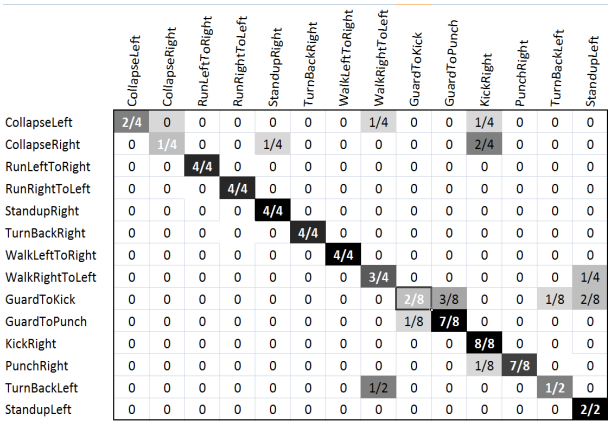


Figure 8. Confusion matrix for novel actor evaluation, **14 actions**

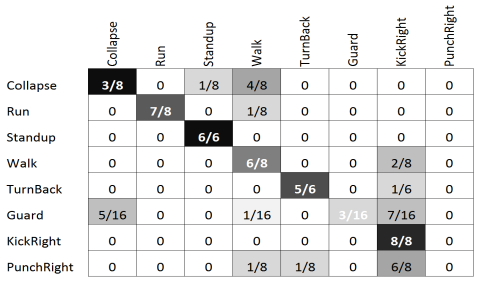


Figure 9. Confusion matrix for novel camera evaluation **8 actions**

Table 5. Accuracy(%) Novel camera view, **8 actions**

	Accurices	
Action class	[13]	Our
Collapse	5/8	5/8
Run	7/8	8/8
Standup	5/6	6/6
Walk	6/8	6/8
TurnBack	5/6	5/6
Guard	3/16	3/16
Kick Right	7/8	8/8
Punch Right	6/8	0/8
Average AccuracyAccrues	50 %	**55.8%**

Table 6. Accuracy(%) Novel camera view, **14 actions**

	Accuracies	
Action class	[13]	Our
Collapse Left	0/4	0/4
Collapse Right	2/4	0/4
Run Left To Right	0/4	3/4
Run Right To Left	0/4	4/4
Stand up Right	4/4	2/4
Stand up Left	2/4	0/2
Walk Left To Right	3/4	4/4
Walk Right To Left	0/4	3/4
Turn Back Right	3/4	2/4
Turn Back Left	2/2	0/2
Guard To Kick	0/8	0/8
Guard To Punch	2/8	0/8
Kick Right	7/8	8/8
Punch Right	6/8	0/8
Average Accuracies	**42.6%**	38.2%

action recognition and will be addressed in future work.

5.4. Evaluation on the Weizmann Data

We also evaluated our approach on the Weizmann data set and considered the 10 contained actions in a leave-one-out cross evaluation. In particular, we experimented with temporal pose templates using MHI where we varied the duration parameter values $\tau = [2, 3, 5, 7]$. This resulted in a 100% classification accuracy for $\tau = 7$. Table 7 compares the results obtained from our approach to established state-of-the-art approaches.

This result conveys a message similar to what is reported in by [12] which found that 1–7 frames are sufficient for basic action recognition. In our case, we found that setting the duration history to $\tau \geq 5$ in constructing MHI temporal templates provides reliable recognition rates in classifying human action.

Table 7. Weizmann data set

	Accuracies		
Approach	Input	Seq	Result (%)
Snippet 1 [12]	Image data	83	93.5%
Snippet 10 [12]	-	-	99.6%
Snippet all seq. [12]	-	-	100%
[5]	Silhouette	90	98.9%
[17]	Silhouette	90	97.8%
[4]	Silhouette	81	100%
[16]	Silhouette	93	96.8%
our approach	Silhouette	93	
$(MHI)_{\tau=2}$	-	-	50.5%
$(MHI)_{\tau=3}$	-	-	86.8 %
$(MHI)_{\tau=5}$	-	-	**98.9%**
$(MHI)_{\tau=7}$	-	-	**100%**

6. Conclusion

In this paper, we reported an extensive evaluation of an approach to human action recognition that models activity sequence by means of groups of temporal templates. In order to extract meaningful and discriminative temporal pose templates for different actions, we perform k-means clustering on pose templates and extract key pose templates for subsequent classification. Our approach is conceptually simple, computationally efficient to implement either in on-line or off-line settings, and robust against variations due to reasonable perspective variations, different actions, and different actors. The proposed approach using Motion History Images (MHIs) was found to perform favorably in comparison to a number of recently proposed state-of-the-art methods. However it inherits a known shortcoming of MHIs which limits its recognition rate in case of severely changing camera views.

References

[1] Md. Ahad, J. Tan, H. Kim, and S. Ishikawa. Motion history image: its variants and applications. *Machine Vision and Applications*, pages 1–27, Oct. 2010.

[2] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE Trans. Pattern recognition and Machine Intelligence*, 23(1):257–267, 2001.

[3] P. Hsiao, C. Chen, and L. Chang. Human action recognition using temporal-state shape contexts. In *Proc. ICPR*, 2008.

[4] N. Ikizler and P. Duygulu. Human action recognition using distribution of oriented rectangular patches. In *ICCV Workshop on Human Motion*, 2007.

[5] V. Kellokumpu, G. Zhao, and M. Pietikinen. Human activity recognition using a dynamic texture based method. In *Proc. BMVC*, 2008.

[6] I. Laptev and T. Lindeberg. Space-time interest points. In *Proc. ICCV*, 2003.

[7] F. Martinez-Contreras, C. Orrite-Urunuela, E. Herrero-Jaraba, H. Ragheb, and S.A. Velastin. Recognizing human actions using silhouette-based hmm. In *Proc. AVSS*, 2009.

[8] S. Mukherjee, S. Biswas, and D. Mukherjee. Human action recognition in video by 'meaningful' poses. In *Proc. ICVGIP*, 2010.

[9] J. Niebles and L. Fei-fei. A hierarchical model of shape and appearance for human action classification. In *Proc.CVPR*, 2007.

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

[11] C. Rao and M. Shah. View-invariant representation and recognition of actions. In *Proc. ICPR*, 2002.

[12] K. Schindler and L. van Gool. Action snippets: How many frames does human action recognition require? In *Proc. CVPR*, 2008.

[13] S. Singh, S.A. Velastin, and H. Ragheb. Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In *Proc. AVSS*, 2010.

[14] X. Sun, M. Chen, and A. Hauptmann. Action recognition via local descriptors and holistic features. In *Proc. CVPR*, 2009.

[15] C. Thurau and V. Hlavac. n-grams of action primitives for recognizing human behavior. In *Proc. CAIP*, 2007.

[16] L. Wang and C. Leckie. Encoding actions via quantized vocabulary of averaged silhouettes. In *Proc. ICPR*, 2010.

[17] L. Wang and D. Suter. Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. In *Proc. CVPR*, 2007.

[18] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2–3):249–257, 2006.

[19] D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, 115(2):224–241, 2011.

[20] Y. Yacoob and M.J. Black. Parameterized modeling and recognition of activities. In *Proc. CVPR*, 1998.