

3D Activity Recognition using Motion History and Binary Shape Templates

Anonymous ACCV 2014 submission

Paper ID 749

Abstract. This paper presents our work on activity recognition in 3D depth images. We propose a global descriptor that is accurate, compact and easy to compute as compared to the state-of-the-art for characterizing depth sequences. Activity enactment video is divided into temporally overlapping blocks. Each block (set of image frames) is used to generate Motion History Templates (MHTs) and Binary Shape Templates (BSTs) over three different views - front, side and top. The three views are obtained by projecting each video frame onto three mutually orthogonal Cartesian planes. MHTs are assembled by stacking the difference of consecutive frame projections in a weighted manner separately for each view. Histograms of oriented gradients are computed and concatenated to represent the motion content. Shape information is obtained through a similar gradient analysis over BSTs. These templates are built by overlaying all the body silhouettes in a block, separately for each view. To effectively trace shape-growth, BSTs are built additively along the blocks.

Consequently, the complete ensemble of gradient features carries both 3D shape and motion information to effectively model the dynamics of an articulated body movement. Experimental results on 4 standard depth databases (MSR 3D Hand Gesture, MSR Action, Action-Pairs, and UT-Kinect) prove the efficacy as well as the generality of our compact descriptor. Further, we successfully demonstrate the robustness of our approach to (impulsive) noise and occlusion errors that commonly affect depth data.

1 Introduction

Action recognition from video sequences is a widely explored challenge, which continues to be the subject of active research. With important applications in the areas of video surveillance, video indexing/retrieval, human-computer interaction (in particular for gaming consoles or surgical assistance), robotic navigation and many others, the task of human activity recognition has immense practical value. Research in this field began on RGB video sequences captured by single traditional cameras – proven successful approaches to static image analysis were extrapolated to cater to the additional time axis. Such techniques can be typically categorized into local descriptor [1], [2], [3] and global descriptor-based [4], [5], [6] approaches.

045 More recently, the advent of affordable depth cameras has shifted the focus 045
046 onto depth images, which can now be easily recorded in real-time with good 046
047 accuracies. Besides providing clear advantages in terms of illumination invariance 047
048 and robust foreground extraction, range images also encode shape and motion 048
049 information which (if represented efficiently) can facilitate much improved ac- 049
050 tivity recognition results. 050

051 Depth images, however, need to be treated differently from their colored 051
052 counterparts. Any attempt to use local differential operators tends to fail, due 052
053 to false firing on discontinuous black regions of undefined depth values. As con- 053
054 firmed in [7], for MSR Daily Activity Dataset [8] 60% of the identified Dollar 054
055 interest points [2] were empirically observed at locations irrelevant to the action . 055
056 Thus, interest-point based 2D video analysis approaches like STIP [1] have been 056
057 shown inadequate to deal with 3D depth data. 057

058 The development of robust approaches to the quick and accurate estimation 058
059 of 3D joint positions of human body from a single depth image [9] have spurred 059
060 substantial research work, aimed at achieving action recognition through skele- 060
061 ton tracking [8,10,11,12]. [8], for instance, proposes the use of Fourier pyramids 061
062 to represent the temporal dynamics of 3D joint-positions, learned using discrim- 062
063 inative actionlet pools and a multiple kernel based approach. Li Xia et al. [13] 063
064 also employ 3D skeletal joint positions for action recognition. Although these ap- 064
065 proaches do yield good practical accuracies, 3D joint positions are prone to noise 065
066 and even partial occlusion can drastically affect recognition. Also, until recently, 066
067 due to the difficulty of efficient skeleton tracking for hands these approaches 067
068 were not employable in hand gesture recognition. Although the approach in [14] 068
069 has overcome such a limitation, the additional tracking step is still susceptible 069
070 to occlusion problems. 070

071 Approaches which do not make use of skeleton tracking have been proposed. 071
072 In [15] an 'action graph' based on bag of 3D contour points is used to iden- 072
073 tify action classes. Methods based on dense sampling have also been attempted, 073
074 which perform a grid-based analysis of the 4D spatio-temporal volume [7] associ- 074
075 ated with a depth sequence for recognizing actions [16]. Although such methods 075
076 can be improved by introducing random sampling of sub-volumes [17] to re- 076
077 duce computation cost and make the features more discriminative, they are still 077
078 computationally demanding, both during training and testing. 078

079 Bobick & Davis [18] introduced the concepts of Motion Energy Templates 079
080 (METs) and Motion History Templates (MHTs) in traditional videos. They rec- 080
081 ognized body movements in input RGB sequences using Hu-moments based sta- 081
082 tistical models built from above templates. Subsequently, Davis [19] improved on 082
083 the approach by using an MHT pyramid to account for varying action-enactment 083
084 speeds, and characterized the resulting motion field using a polar histogram. Tian 084
085 et al. [20] also employed MHTs for action recognition in crowded scene videos, 085
086 owing to its effectiveness in capturing continuous and recent foreground motion 086
087 even in a cluttered and inconsistently moving background. Using Harris corner 087
088 points to eliminate noisy motion from MHT, they computed gradient features 088
089 from intensity image and MHT. The features were learnt by a Gaussian Mix- 089

090 ture Model (GMM) based classifier. However, it was Yang et al. [21] who first 090
091 extended the idea to 3D depth images. They adopted METs and stacked motion 091
092 regions for entire video sequences over three distinct views – front, side and 092
093 top. Gradient features for the 3 consolidated binary templates were then used 093
094 to represent depth sequences. 094

095 095

096 **1.1 Contribution** 096

097 097

098 Based on the conducted critical review of past approaches and their limitations 098
099 we propose a highly discriminative ensemble of both shape and motion features, 099
100 extracted over a novel temporal construct formed by overlapping video blocks, 100
101 for 3D action recognition from depth sequences. 101

102 First of all, contrarily to [21] where a single template is built over the complete 102
103 video sequence, we build and analyze motion and shape templates separately for 103
104 overlapping temporal blocks. The intuition behind this novel temporal construct 104
105 and its discriminative advantages are elaborated in 3.1. By dividing the video 105
106 sequence in smaller temporal blocks, we prevent the loss of motion information 106
107 which happens when a more recent action overwrites an old action at the same 107
108 point. Temporal overlap across blocks maintains continuity in the templates. 108

109 Secondly, instead of using Depth Motion Maps (a form of METs) as in [21], 109
110 we adopt Motion History Templates so as to effectively capture information on 110
111 the direction of motion, in the form of gray level variations. This allows our 111
112 approach to cope successfully with sequences like those in the 3D Action Pairs 112
113 database [7], in which the direction of motion is the essential discriminating 113
114 factor. 114

115 In an additional original feature of our proposal, METs are assembled over 115
116 temporal blocks in an accumulative manner to capture shape information. The 116
117 shape of the resulting black and white boundaries describes how the human body 117
118 envelope grows in the course of a certain action. 118

119 Finally, histograms of oriented gradients extracted from these templates are 119
120 concatenated to yield the final feature vector. The latter constitutes a novel, 120
121 low-dimensional, computationally inexpensive and highly discriminative global 121
122 descriptor for depth sequences. This descriptor is used to train a non-linear 122
123 large-margin classifier (RBF kernel based SVM) to identify action classes. The 123
124 complete process flow is shown in Fig.1. 124

125 The generality of the approach is empirically demonstrated by evaluating it 125
126 on 4 standard depth benchmarks – the 3D Hand Gesture, 3D Action, Action Pair, 126
127 and UT-Kinect datasets. A vanilla implementation of our method outperforms 127
128 the existing-best on 2 out of three 3 standard benchmarks, and has performance 128
129 equal to the state of the art on the skeleton-biased UT-Kinect dataset. Note 129
130 that our approach uses a block-size and an overlap-factor as parameters, that 130
131 are in these experiments still manually selected by trial & error. By setting these 131
132 parameters via cross validation our work has much scope for further performance 132
133 gains. 133

134 Last but not least, in tests commonly neglected in the literature, we exper- 134
135 imentally demonstrate the robustness of our approach to noise and occlusions

that typically affect depth images by artificially corrupting sequences from the MSR 3D Action dataset. Results show that performance is remarkably robust to both nuisances, making our approach suitable for real-world deployment.

1.2 Paper outline

The paper is organized as follows. Section 2 compares our approach in detail with the closest related work. Section 3 illustrates the various steps of the proposed approach. Section 4 recalls the benchmark datasets considered in our tests and details the experimental settings. The performance of our algorithm on the four considered benchmarks is illustrated in Section 5. Section 6 studies the robustness of the approach against common occlusion and noise errors. Section 7 concludes the paper with a word about the future directions of research.

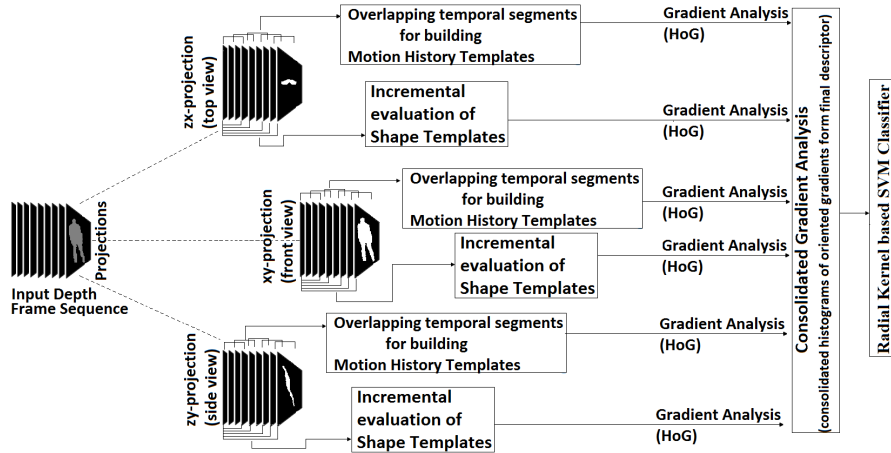


Fig. 1: Process flow for the proposed 3D activity recognition approach.

2 Relevant Related Work

Vennila et al.[22] follow a similar idea to that proposed in our approach, but in a rather more simplistic way. Motion and 3D shape information are not considered in conjunction. Their analysis of only frontal MHIs leaves out depth (3D shape) information, while their Average Depth and Difference Depth Images do not consider motion and thus capture background-clutter depth details also.

Weinland et al. [23] suggest the use of motion history volumes for analyzing voxelset sequences. Their normalization along the r & z axes may cause alignment failure, like confusion between a person spreading his arms and a person dropping them. The Fourier analysis that they apply in concentric circles at

180 each height does not capture limb movement as well as our gradient analysis 180
181 of MHTs. For composite actions such as the walking gait, only one temporal 181
182 segment is used. All these cause confusion between well distinct actions such as 182
183 sit-down & pick-up, turn-around & majority-of-other-actions, walk & pick-up or 183
184 kick. Also, the used 1D-FT overlooks motion symmetry around z axis and may 184
185 fail to distinguish between single & two arm waves, forward & side punch, and 185
186 forward & side kick. 186

187 Finally, Oreifej and Liu [7] represent the current state-of-the-art for the task 187
188 at hand. Their method computes and concatenates the histograms of oriented 188
189 surface normals for each grid cell and uses the final feature descriptor to train 189
190 an SVM classifier. We compare ourselves to [7] empirically in Section 5. 190

191 3 Proposed Approach 191

192 3 Proposed Approach 192

193 3 Proposed Approach 193

194 3 Proposed Approach 194

195 In a depth image, each pixel contains a value which is a function of the depth/distance 195
196 at which the camera encounters an object at the corresponding point in the real 196
197 scene. For a depth camera recording from the front, the information of overlapped 197
198 object points in the rear is not captured. Therefore, a depth image at 198
199 any time does not contain an exact all-view model of the 3D objects present 199
200 in the scene. Nonetheless, the data is sufficient to generate three approximate 200
201 front, side and top views of the scene which, for a single 3D object, contain im- 201
202 portant, non-redundant and discriminative information both in terms of shape 202
203 (over a single frame) and motion (over a sequence of frames). It is to obtain and 203
204 leverage this enhanced 3D shape and motion information that we first project 204
205 each frame of the incoming video sequence onto the three orthogonal Cartesian 205
206 planes. The front, side and top views so obtained for a given frame are as il- 206
207 lustrated in Fig.2. Construction of motion and shape information from these 207
208 projections is illustrated in Fig.3, and described in the remainder of the Section. 208

209 3.1 Motion Information Evaluation 209

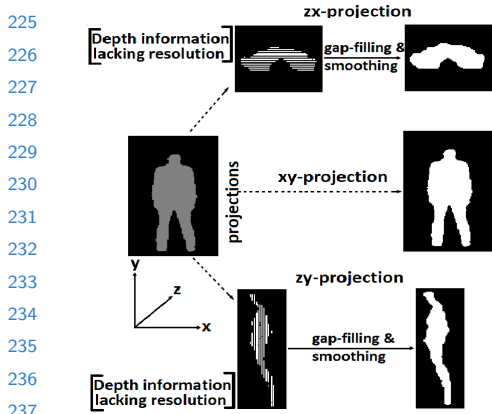
210 3.1 Motion Information Evaluation 210

211 3.1 Motion Information Evaluation 211

212 We split the input video sequence into overlapping temporal blocks. The split 212
213 allows motion to be captured over a smaller number of frames thus enabling a 213
214 more detailed analysis of motion. As seen in Fig.3(Set 1), if one motion history 214
215 template is built for the complete action of 'draw tick', any old motion informa- 215
216 tion gets overwritten when a more recent action occurs at the same point. Not 216
217 only is part of the motion information lost, but the exact direction of motion 217
218 (represented by the direction of increasing gray intensity) also gets distorted. 218

219 Using overlapping temporal blocks helps to ensure a smooth connectivity of 219
220 the articulated motion across consecutive blocks. Note that the movement is 220
221 more continuous in Fig.3(Set 3) as compared to Fig.3(Set 2), in which such a 221
222 temporal overlap is not maintained. 222

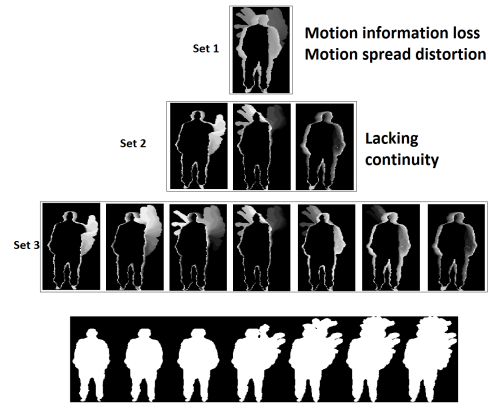
223 As mentioned before, the three orthogonal views are handled separately. 223
224 Firstly, for the complete video sequence we gather all the frame projections 224



225
226
227
228
229
230
231
232
233
234
235
236
237

Fig. 2: Depth images projected onto three Cartesian planes to obtain front, side and top views. (Discontinuities resolved using gap-filling and smoothing.)

238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269



225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269

Fig. 3: (Top) Motion history templates for (Set 1) complete sequence, (Set 2) non-overlapping blocks, (Set 3) overlapping blocks; (Bottom) Binary Shape Templates for a 'draw tick' action focused in the top-right region.

for each view. Then, separately for each view, we divide the frames into overlapping blocks and proceed to build their motion field. For each block, consecutive frames are subtracted to identify the parts in motion. These parts are stacked together in a weighted manner, with the weight being proportional to the recency of motion, to obtain the desired Motion History Templates.

Namely, the following operations are performed on each block.

A motion template MT_f , the difference between two consecutive body silhouettes B_f , is computed as: $MT_f = B_{f+1} - B_f$ for each frame $f \in 1, \dots, n - 1$ in the block. The motion history template MHT_b for block b is computed iteratively, by first initializing it to null, and then updating it by applying for each $f \in 1, \dots, n - 1$ a pixel-wise maximum ($pmax$) of the current motion history template and the weighted motion template $f * MT_f$, namely: $MHT'_b = pmax(MHT_b, f * MT_f)$.

3.2 Shape Information Evaluation

Shape information is constructed in the form of gradient features of binary shape templates, compiled over temporal blocks of the input video sequence. As previously mentioned the three orthogonal views are processed separately. Frame projections for each view are collected and divided into overlapping blocks for which shape templates are constructed. Shape templates become clearer over longer intervals of time and are thus incrementally built, as shown in Fig.3(bottom). The shape template for block 2 is added over the shape template for block 1, the

270 template for block 3 is built over the shape templates for blocks 1 and 2, and 270
271 so-on. 271

272 Thus, the shape template for a given block is built by additive overlaying 272
273 of body silhouettes, via a frame on frame pixel-wise OR operation, for all the 273
274 frames in the current as well as the previous blocks. 274

275 First, the shape template is initialized to $ST_b = ST_{b-1}$, for $b > 1$ (a matrix of 275
276 zeros otherwise). Then, ST_b is updated iteratively via a pixel-wise OR of the 276
277 current shape template and the current frame B_f for all frames $f \in 1, \dots, n$ as: 277
278 $ST'_b = ST_b || B_f$. 278

279 279

280 **3.3 Gradient Analysis** 280

281 281

282 Motion history is contained in gray-level variations (MHTs) while shape is rep- 282
283 resented by white-object boundaries (BSTs). Extraction of shape and motion 283
284 information therefore necessitates a form of gradient analysis. Herein lies our 284
285 motivation behind the use of histograms of oriented gradients [24] to obtain the 285
286 final feature descriptor for the input depth sequence. 286

287 For the 3D Hand Gesture Dataset, a single histogram of gradients is com- 287
288 puted for a complete template, i.e, the standard implementation, where his- 288
289 tograms are built over sets of grid cells and later concatenated, is not followed. 289
290 We observe that different hand gesture enactments have different proportions of 290
291 gradient values. Thus, without the need for greater discrimination, positions of 291
292 gradient values could be bypassed. 292

293 For whole-body datasets, however, the standard HoG implementation is fol- 293
294 lowed. Different regions or body parts can make the same gradient contribution. 294
295 For examples, actions pairs like - “side boxing and side kick”, “forward punch and 295
296 forward kick”, and “high arm wave and two hand wave”; show similar gradient 296
297 proportions over the complete image but the contributing gradient regions are 297
298 different. Thus, for a better discrimination between different action enactments 298
299 a region-wise gradient analysis is essential. 299

300 300

301 **4 Experimental Settings** 301

302 302

303 **4.1 Datasets** 303

304 304

305 We have evaluated the performance of our proposed approach on 4 standard 305
306 depth datasets - MSR 3D Gesture Dataset [25], MSR 3D Action Dataset [15], 3D 306
307 Action Pair Dataset [7], and UT-Kinect Database[13]. Details of each database, 307
308 their preprocessing operations, and the parameter settings for feature extraction 308
309 are as elaborated below. 309

310 **MSR 3D Gesture Dataset:** The Gesture3D dataset [25] is a hand gesture 310
311 dataset of depth sequences captured using a depth camera (more particularly a 311
312 Kinect device). It contains a set of 12 dynamic American Sign Language (ASL) 312
313 gestures, namely - “bathroom”, “blue”, “finish”, “green”, “hungry”, “milk”, 313
314 “past”, “pig”, “store”, “where”, “j”, “z”. 10 subjects have performed each hand 314

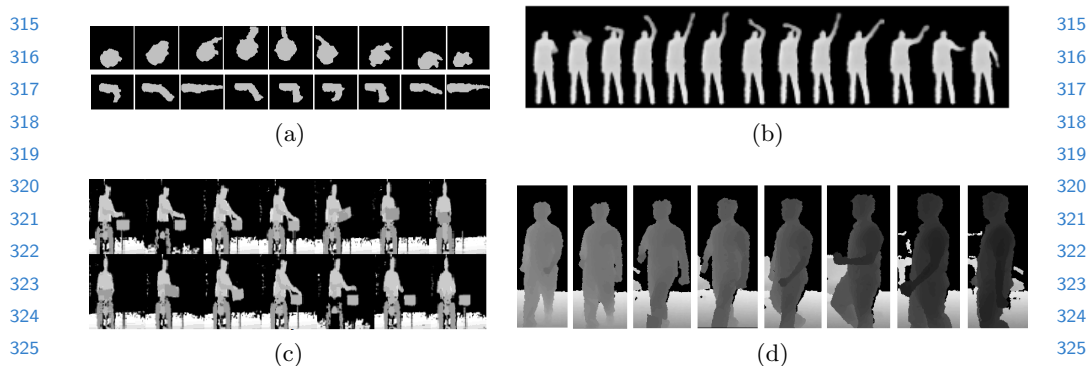


Fig. 4: Sample frames for (a) Hand gesture enactments of 'where', 'pig' (b) Action category of 'high arm wave' (c) Action pairs of 'pick up' & 'put down a box' (d) UT Kinect action category of 'walk'.

gesture 2-3 times yielding a total of 336 sample sequences. Sample frames for two gesture categories are shown in Fig.4a. The dataset contains images of hand portions segmented above the wrist and is highly challenging due to self-occlusions in gesture enactments.

MSR 3D Action Dataset: MSR Action3D dataset [15] is an action dataset of depth sequences captured using a depth camera similar to the Kinect device. This dataset contains twenty action categories: “high arm wave”, “horizontal arm wave”, “hammer”, “hand catch”, “forward punch”, “high throw”, “draw x”, “draw tick”, “draw circle”, “hand clap”, “two hand wave”, “side-boxing”, “bend”, “forward kick”, “side kick”, “jogging”, “tennis swing”, “tennis serve”, “golf swing”, “pickup & throw”. Each action is performed around 2-3 times by 10 subjects, yielding a total of 567 action samples. Subjects face the camera in the depth maps which are captured at a rate of 15 fps and a resolution of 320×240 . Sample frames for a 'high arm wave' action are shown in Fig.4b.

The 20 actions were chosen in the context of using the actions to interact with gaming consoles. They reasonably cover various movements of arms, legs, torso and their combinations. In addition, for actions performed by a single arm or leg, the subjects were advised to use their right arm or leg. In this dataset, the background has been cleaned to remove the discontinuities due to undefined depth points. However, recognition remains highly challenging due to close similarities between different actions.

3D Action Pair Database: This database was introduced by Oreifej and Liu [7] to emphasize the point that two different actions may have similar shape and motion cues, but the correlation between these cues may vary. Quoting the example from [7], “Pick up” and “Put down” actions have similar motion and shape; however, the co-occurrence of object-shape and hand-motion is in different spatio-temporal order and that holds the key to distinguishing the two. Based on this idea, the database puts together 6 action pairs for classification,

360 namely - “Pick up a box/Put down a box”, “Lift a box/Place a box”, “Push a
361 chair/Pull a chair”, “Wear a hat/Take off a hat”, “Put on a backpack/Take off a
362 backpack”, “Stick a poster/Remove a poster”. Each action is performed 3 times
363 using 10 different actors yielding a total of 360 depth sequences. Sample frames
364 presenting an action-pair are as shown in Fig.4c. 364

365 **UT Kinect Action Database:** This database was compiled by Li Xia et al. 365
366 using a single stationary Kinect camera. It includes 10 action types: “walk”, “sit-
367 down”, “stand-up”, “pick-up”, “carry”, “throw”, “push”, “pull”, “wave-hands”,
368 “clap-hands”. Each of the 10 subjects perform each action twice. The UT-Kinect
369 database is highly challenging, with significant variations in single action enact-
370 ment (‘Throw’ with left /right arm, ‘Carry’ with one/both hands, ‘walk’ - first
371 front-ward and then side-ward as in Fig.4d), action clips duration, frequent oc-
372 clusions due to changing views, and an annoying shifting background clutter
373 possibly due to camera trembling. 373

374 With the frames being recorded only when a skeleton is tracked, the database
375 is biased towards skeleton based processing. The above frame selection also re-
376 duces the practical frame rate from 30fps to 15fps. Furthermore, around 2%
377 of the frames have multiple skeletons recorded with slightly different joint loca-
378 tions. Again, using skeleton information the main object can be easily segmented,
379 while in general depth-information processing (as adopted by us) this introduces
380 additional challenges. 380

381 4.2 Preprocessing 382

383 Images in the hand gesture dataset have varying sizes. For experimental pur-
384 poses, the images are used as is without any cropping and/or resizing. The
385 aim is to preserve the original shape and gradients in the image. Values in the
386 histograms of gradients are normalized and carry information of the relative gra-
387 dient proportion. Thus, image size does not affect this normalized distribution. 388

389 For the 3 full-body datasets (MSR Action, Action-Pair and UT-Kinect) im-
390 ages need to be uniformly sized to facilitate a more effective grid-based HoG
391 analysis. Also, we can observe that depth information is discontinuous. As can
392 be seen in Fig.2, depth values appear in steps resulting in the black slit-like gaps.
393 The gaps introduce irrelevant gradients. To remove the gaps, we divide each
394 depth value by the average step-size (for gap-filling) and perform small-sized
395 structuring element-based erosion and dilation (for smoothing). This is followed
396 by image packing and resizing (if required). Finally, the front-view, side-view and
397 top-view templates for the 3 databases are sized: (240x160, 240x180, 180x160);
398 (240x160, 240x480, 480x160); and (320x240, 320x240, 320x320) respectively. 399

400 4.3 Parameter settings 401

402 To compute our motion analysis descriptors, video sequences (say, of f frames) 401
403 are divided into t units along the time axis. We move temporally in size of a 402
404 unit (f/t frames) and in steps decided by the overlap factor (o), to assemble the 403
405 blocks along the video sequence. Essentially, each block contains f/t frames, and 404

405 shares $1/o$ amount of overlap with the previous block. Alternatively, this can be 405
406 seen as dividing the video into $t * o$ frame sets and grouping o such sets starting 406
407 at each, to yield a total of: $b = (t * o) - (o - 1)$ blocks. 407

408
409 For each block, motion as well as binary shape templates are constructed for 409
410 each of the 3 orthogonal views. Thus, for any given video sequence, a maximum 410
411 of $(3*2*b)$ templates are analyzed. For motion history templates, the direction of 411
412 motion is stored in grey level variations: hence, all gradients in range $0-360^\circ$ are 412
413 considered. For our HoG-based analysis, this range is divided into b_1 gradient 413
414 bins. For shape analysis, a 90° spread was found to be adequate and the $0-90^\circ$ 414
415 range is divided into b_2 gradient bins. Finally, all the histograms are normalized 415
416 using the classical L2-norm [24] and concatenated to form the overall feature 416
417 descriptor. 417

418
419 Dataset specific values for all these parameters are provided below. Presently 419
420 these are empirically decided, however subsequently we aim to set these param- 420
421 eters via cross validation to boost our performance. 421

422
423 **MSR 3D Gesture Dataset:** With $t = 3$ and $o = 3$, the number of blocks b 423
424 $= 7$. We build and analyze a total of 42 $(3*2*7)$ templates to extract motion and 424
425 shape information. As mentioned previously, for hand gesture templates gradient 425
426 analysis is done once for the complete image without any division into smaller 426
427 grids. b_1 and b_2 are set to 72 and 18 respectively. 427

428
429 **MSR 3D Action Dataset:** With $t = 3$ and $o = 3$, the number of blocks 429
430 $b = 7$. We build and analyze a total of 42 $(3*2*7)$ templates. The 7 motion 430
431 history templates and 7 shape templates for front-view of 'draw tick' action are 431
432 as shown in Fig.3 (Set 3) & (bottom). Further, for HoG analysis, the front, side 432
433 and top views of motion history templates are divided into $6x4$, $6x4$ and $4x4$ 433
434 non-overlapping grid blocks while for shape templates they are divided into $1x1$, 434
435 $1x1$ and $1x1$ sized grid blocks. b_1 and b_2 are set to 36 and 9 respectively. 435

436
437 **3D Action Pair Database:** With $t = 4$ and $o = 2$, the number of blocks 437
438 $b = 7$. We build and analyze a total of 42 $(3*2*7)$ templates. For HoG analysis, 438
439 front, side and top views of motion history templates are divided into $6x4$, $6x8$ 439
440 and $8x4$ non-overlapping grid blocks while for shape templates they are divided 440
441 into $1x1$, $1x1$ and $1x1$ sized grid blocks. b_1 and b_2 are set to 36 and 9 respectively. 441

442
443 **UT Kinect Action Database:** With $t = 4$ and $o = 2$, the number of blocks 442
444 $b = 7$. We build and analyze a total of 21 $(3*1*7)$ templates. Due to significant 443
445 variations in view and action enactment, as well frequent occlusions and camera 444
446 motion, we observed shape templates to be counter-productive and hence shape 445
447 templates were not considered. For HoG analysis, front, side and top views of 446
448 motion history templates are divided into $3x4$, $3x4$ and $4x4$ non-overlapping grid 447
448 blocks. In place of shape templates, motion history templates are additionally 448
449 analyzed over $1x1$, $1x1$ and $1x1$ sized grid blocks. b_1 is set to 36. 449

Approach	Accuracy (%)
Proposed (MHI + BST based Gradient Analysis)	96.6
Oreifej & Liu[7]	92.45
Jiang et al.[17]	88.5
Kurakin et al.[25]	87.77

Fig. 5: Accuracy comparison for 3D Hand Gesture Recognition.

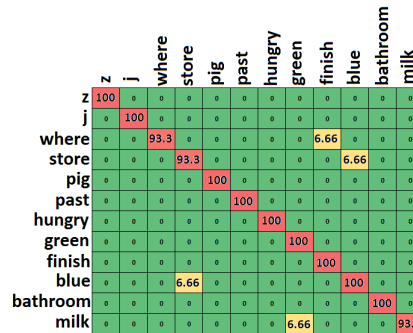


Fig. 6: Confusion matrix for 3D hand gesture recognition using the proposed approach.

5 Recognition Results

We use SVM [26] discriminative classifier in one-versus-one configuration for the task of multi-class classification. In order to handle non-linear class boundaries SVM uses radial basis function kernel.

MSR 3D Gesture Dataset: As per the standard protocol, for every gesture category samples of first 5 subjects form the training set while those of next 5 subjects form the test set. With this cross-subject accuracy of 96.6% our proposed method significantly outperforms the existing best [7]. Fig.5 lists the comparison with the previously attempted approaches for the given dataset. The related confusion matrix is presented in Fig.6.

MSR 3D Action Dataset: Our approach yields a cross-subject accuracy of 83.8%. Fig.7 presents the performance comparison with previously attempted approaches for the given dataset. Related confusion matrix is presented in Fig.8. Jiang et al. [8] obtain an accuracy figure of 88.20% with skeleton tracking approach. However, it has the drawback of being dependent on the precision of 3D joints positions making it susceptible to occlusion errors. Yang et al. [21], Vieira et al.[16], Li et al. [15] perform experiments on three subsets of the 3D Action dataset – AS1, AS2 and AS3, containing 8 action categories each. Confusion among a set of categories is always greater than or equal to the confusion in its subsets. Hence, the correct overall-accuracy for the approaches is not the average of the 3 subset accuracies but lesser than the minimum amongst the 3 values i.e. < 84.1%, < 81.30% and < 71.90% respectively. Considering the HoN4D approach on equal grounds i.e. without data-specific quantization refining [7], the accuracy of our approach is 2% lower.

Given the performance of our approach on the other 3 datasets, we can argue that the relatively low accuracy is mainly due to inadequate data. The number of classes is the highest (20), but the number of sample videos per class is the same as in other datasets. We believe that with a higher number of videos our approach would generalize better to test set.

Approach	Accuracy (%)
Proposed (MHI + BST based Gradient Analysis)	83.8
Oreifej & Liu [7]	88.89
Jiang et al. [8]	88.2
Jiang et al. [17]	86.5
Yang et al. [21]/ [7]	<84.1/ 85.52
Vieira et al.[16]	<81.30
Li et al. [15]	<71.90

Fig. 7: Accuracy comparison for 3D Action Data Recognition.

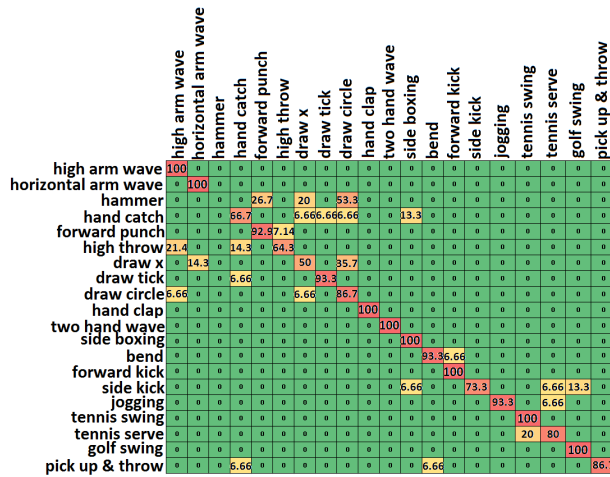


Fig. 8: Confusion matrix for 3D action data recognition using the proposed approach.

3D Action Pair Database: A cross-subject accuracy of 97.22% indicates that if sufficient amount of data is available per action category, our proposed approach has the potential to give top recognition results. Fig.9 presents the performance comparison with other approaches. The related confusion matrix is shown in Fig.10.

UT Kinect Action Database: To the best of our knowledge, our approach is the only non-skeleton based technique attempted for UT Kinect Database. Despite the data being biased for skeletal analysis, as explained in Section 4.1, our method gives a competitive accuracy of 90% (under the same testing conditions as in [13]) compared to the authors' 90.92%.

Approach	Accuracy (%)
Proposed (MHI + BST based Gradient Analysis)	97.22
Oreifej & Liu [7]	96.67
Jiang et al. [8]	82.22
Yang et al. [21]	66.11

Fig. 9: Accuracy comparison for 3D action-pair data recognition.

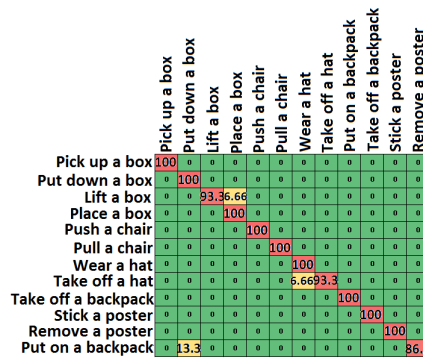


Fig. 10: Confusion matrix for 3D action-pair data recognition using the proposed approach.

6 Performance under Occlusion and Noise

The performance of the proposed approach has also been evaluated in the presence of two noise effects as elaborated in the subsections below. Results have been compiled on MSR 3D Action Database.

6.1 Occlusion

To analyze and compare the robustness of our method in the presence of occlusion, we carried out an experiment similar to that presented in [15]. As shown in Fig.11, for a given occlusion setting, a single quadrant or a combination of 2 quadrants may be occluded in the depth image. Occlusion is incorporated in the test set and Fig.12 shows the relative accuracy for each case. (Relative accuracy is defined as the recognition accuracy with occlusion, as a percentage of the accuracy without any occlusion).

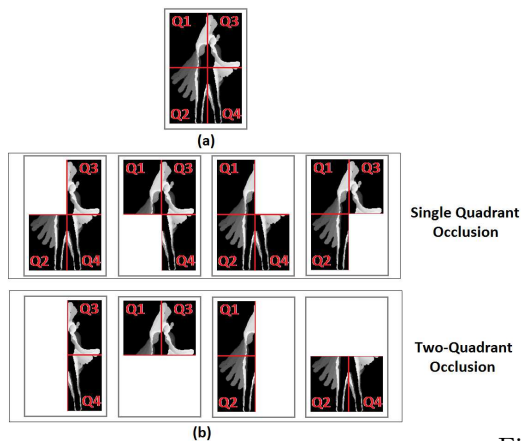


Fig. 11: Occlusion of quadrant(s) for robustness testing.

Occluded Quadrant(s)	Relative Accuracy (%)
none	100
Q1	91.57
Q2	95.99
Q3	64.66
Q4	96.39
Q1+Q3	43.376
Q3+Q4	62.25
Q4+Q2	85.948
Q2+Q1	65.87

Fig. 12: Recognition performance of the proposed approach in presence of occlusion.

With most of the actions focused on the top-right of the image, occlusion of Q3 gives the highest error rate. For all other single quadrant occlusions, accuracy reduction is no greater than 9%. Our method has the potential to handle occlusions primarily because it performs an independent region-wise gradient analysis over motion history and shape templates. Even if a particular part of the image is occluded, visible body regions are unaffected and can be independently analyzed.

6.2 Pepper Noise

The most common error that affects depth images, and consequently renders interest point approaches ineffective, is the presence of discontinuous black re-

585 gions in the images. In a completely new performance measure, we deliberately
586 introduce pepper noise in different percentages (of the total number of image
587 pixels) in the depth images, as in Fig.13. From Fig.14, the recognition accuracy
588 does not fall by more than 6% even in up to 10% pepper noise. This indicates
589 the high robustness of our method to depth discontinuities.

590

591

592

593

594

595

596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

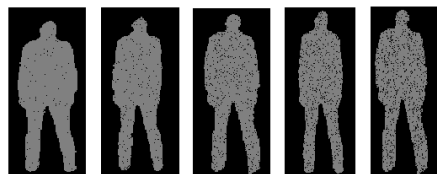


Fig. 13: Images affected by pepper noise in increasing percentages of total no. of pixels - 1%, 2.5%, 5%, 7.5% and 10% .

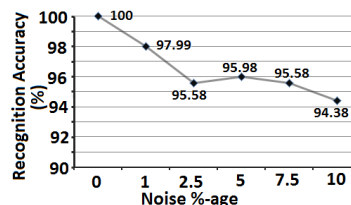


Fig. 14: Recognition performance of the proposed approach in presence of varying amount of pepper noise (to simulate depth discontinuities).

7 Conclusion and Future Work

This paper presents a compact and inexpensive global descriptor for activity recognition from depth videos. The overall approach performs consistently well on 4 very different databases. It outperforms the existing best on 2 out of 3 standard depth databases and is the first non-skeleton based approach to give highly competitive results on UT Kinect Action Database. This proves its generic performance and wide applicability. Also, experimental results confirm its ability to handle occlusion and noise errors that commonly affect depth images.

As part of the future work, we will set the parameters of block-size ($1/t$) and overlap-factor (o) via cross-validation to achieve improved performance through a database-specific setting. Also, we aim to assess the applicability of our approach to real-time action recognition from online depth videos. Temporal blocks can be treated as independent observation states. A model can be made to learn the observation-state-transition-probability to action-category mapping. Thus, temporal blocks and their sequence can help identify the action enacted in a live video. With every incoming block, the decision needs to be reconsidered in the light of the most recent input information.

References

1. Laptev, I., Lindeberg, T.: Space-time interest points. In: ICCV, IEEE Computer Society (2003) 432–439

- 630 2. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse 630
631 spatio-temporal features. In: Proceedings of the 14th International Conference on 631
632 Computer Communications and Networks. ICCCN '05, Washington, DC, USA, 632
633 IEEE Computer Society (2005) 65–72 633
- 634 3. Kläser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 634
635 3d-gradients. In Everingham, M., Needham, C.J., Fraile, R., eds.: BMVC, British 635
636 Machine Vision Association (2008) 636
- 637 4. Yilmaz, A., Shah, M.: Actions sketch: A novel action representation. In: CVPR 637
638 (1), IEEE Computer Society (2005) 984–989 638
- 639 5. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time 639
640 shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **29** (2007) 2247–2253 640
- 641 6. Li, W., Zhang, Z., Liu, Z.: Expandable data-driven graphical modeling of human 641
642 actions based on salient postures. *IEEE Trans. Circuits Syst. Video Techn.* **18** 642
643 (2008) 1499–1510 643
- 644 7. Oreifej, O., Liu, Z.: Hon4d: Histogram of oriented 4d normals for activity recognition 644
645 from depth sequences. In: Computer Vision and Pattern Recognition (CVPR), 645
646 2013 IEEE Conference on. (2013) 716–723 646
- 647 8. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition 647
648 with depth cameras. In: Computer Vision and Pattern Recognition (CVPR), 648
649 2012 IEEE Conference on. (2012) 1290–1297 649
- 650 9. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kip- 650
651 man, A., Blake, A.: Real-time human pose recognition in parts from single depth 651
652 images. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Con- 652
653 ference on. (2011) 1297–1304 653
- 654 10. Yun, H., Sheng-Luen, C., Jeng-Sheng, Y., Qi-Jun, C.: Real-time skeleton-based 654
655 indoor activity recognition. In: Control Conference (CCC), 2013 32nd Chinese. 655
656 (2013) 3965–3970 656
- 657 11. Shuzi, H., Jing, Y., Huan, C.: Human actions segmentation and matching based 657
658 on 3d skeleton model. In: Control Conference (CCC), 2013 32nd Chinese. (2013) 658
659 5877–5882 659
- 660 12. Yu, X., Wu, L., Liu, Q., Zhou, H.: Children tantrum behaviour analysis based 660
661 on kinect sensor. In: Intelligent Visual Surveillance (IVS), 2011 Third Chinese 661
662 Conference on. (2011) 49–52 662
- 663 13. Xia, L., Chen, C.C., Aggarwal, J.: View invariant human action recognition using 663
664 histograms of 3d joints. In: Computer Vision and Pattern Recognition Workshops 664
665 (CVPRW), 2012 IEEE Computer Society Conference on. (2012) 20–27 665
- 666 14. : (Efficient model-based 3d tracking of hand articulations using kinect) 666
- 667 15. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3d points. In: Com- 667
668 puter Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer 668
669 Society Conference on. (2010) 9–14 669
- 670 16. Vieira, A., Nascimento, E., Oliveira, G., Liu, Z., Campos, M.: Stop: Space-time oc- 670
671 cupancy patterns for 3d action recognition from depth map sequences. In Alvarez, 671
672 L., Mejail, M., Gomez, L., Jacobo, J., eds.: Progress in Pattern Recognition, Im- 672
673 age Analysis, Computer Vision, and Applications. Volume 7441 of Lecture Notes 673
674 in Computer Science. Springer Berlin Heidelberg (2012) 252–259 674
- 675 17. Wang, J., Liu, Z., Chorowski, J., Chen, Z., Wu, Y.: Robust 3d action recognition 675
676 with random occupancy patterns. In: Proceedings of the 12th European Conference 676
677 on Computer Vision - Volume Part II. ECCV'12, Berlin, Heidelberg, Springer- 677
678 Verlag (2012) 872–885 678
- 679 18. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal 679
680 templates. *IEEE Trans. Pattern Anal. Mach. Intell.* **23** (2001) 257–267 680

675	19. Davis, J.W.: Hierarchical motion history images for recognizing human motion. In:	675
676	IEEE Workshop on Detection and Recognition of Events in Video. (2001) 39–46	676
677	20. Tian, Y., Cao, L., Liu, Z., Zhang, Z.: Hierarchical filtered motion for action recog-	677
678	nition in crowded videos. Systems, Man, and Cybernetics, Part C: Applications	678
679	and Reviews, IEEE Transactions on 42 (2012) 313–323	679
680	21. Yang, X., Zhang, C., Tian, Y.: Recognizing actions using depth motion maps-based	680
681	histograms of oriented gradients. In: Proceedings of the 20th ACM International	681
682	Conference on Multimedia. MM '12, New York, NY, USA, ACM (2012) 1057–1060	682
683	22. Megavannan, V., Agarwal, B., Venkatesh Babu, R.: Human action recognition	683
684	using depth maps. In: Signal Processing and Communications (SPCOM), 2012	684
685	International Conference on. (2012) 1–5	685
686	23. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using	686
687	motion history volumes. Comput. Vis. Image Underst. 104 (2006) 249–257	687
688	24. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In:	688
689	Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer	689
690	Society Conference on. Volume 1. (2005) 886–893 vol. 1	690
691	25. Kurakin, A., Zhang, Z., Liu, Z.: A real time system for dynamic hand gesture	691
692	recognition with a depth sensor. In: Signal Processing Conference (EUSIPCO),	692
693	2012 Proceedings of the 20th European. (2012) 1975–1979	693
694	26. Chang, C.C., Lin, C.J.: Libsvm: A library for support vector machines. ACM	694
695	Trans. Intell. Syst. Technol. 2 (2011) 27:1–27:27	695
696		696
697		697
698		698
699		699
700		700
701		701
702		702
703		703
704		704
705		705
706		706
707		707
708		708
709		709
710		710
711		711
712		712
713		713
714		714
715		715
716		716
717		717
718		718
719		719