An Overview of Contest on Semantic Description of Human Activities (SDHA) 2010

M. S. Ryoo[1,2], Chia-Chih Chen[1], J. K. Aggarwal[1], and Amit Roy-Chowdhury[3]

[1]Computer and Vision Research Center, the University of Texas at Austin, USA
[2]Robot/Cognition Research Department, ETRI, Korea
[3]Video Computing Group, Dept. of EE, University of California, Riverside, USA
mryoo@etri.re.kr, {ccchen,aggarwaljk}@mail.utexas.edu, amitrc@ee.ucr.edu
http://cvrc.ece.utexas.edu/SDHA2010/

Abstract. This paper summarizes results of the 1st Contest on Semantic Description of Human Activities (SDHA), in conjunction with ICPR 2010. SDHA 2010 consists of three types of challenges, High-level Human Interaction Recognition Challenge, Aerial View Activity Classification Challenge, and Wide-Area Activity Search and Recognition Challenge. The challenges are designed to encourage participants to test existing methodologies and develop new approaches for complex human activity recognition scenarios in realistic environments. We introduce three new public datasets through these challenges, and discuss results of state-of-the-art activity recognition systems designed and implemented by the contestants. A methodology using a spatio-temporal voting [19] successfully classified segmented videos in the UT-Interaction datasets, but had a difficulty correctly localizing activities from continuous videos. Both the method using local features [10] and the HMM based method [18] recognized actions from low-resolution videos (i.e. UT-Tower dataset) successfully. We compare their results in this paper.

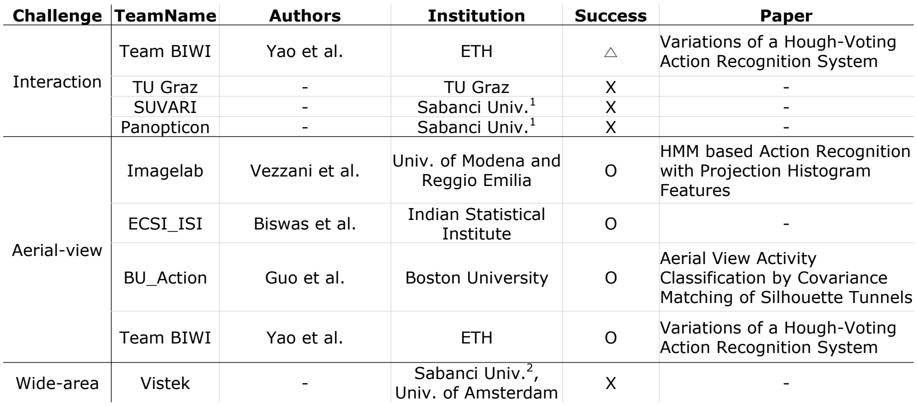
Keywords: Activity recognition contest, human activity recognition, video analysis

1 Introduction

Human activity recognition is an area with an increasing amount of interest, having a variety of potential applications. An automated recognition of human activities from videos is essential for the construction of smart surveillance systems, intelligent robots, human-computer interfaces, quality of life devices (e.g. elderly monitoring), and military systems. Developments of spatio-temporal feature extraction, tracking, and high-level activity analysis are leading today's computer vision researchers to explore human activity recognition methodologies practically applicable for real world applications.

In this contest, we propose three types of activity recognition challenges which focus on different aspects of human activity recognitions: High-level Hu-

Table 1. A table summarizing the results of SDHA 2010 contest. We made the authors of the teams who decided not to submit their results anonymous. We invited three teams who showed the best results to submit their papers [9, 17, 21].



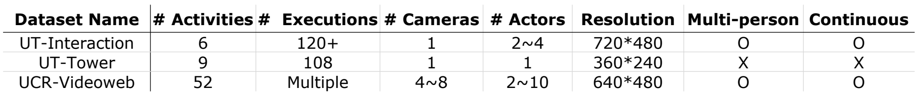
Challenge	TeamName	Authors	Institution	Success	Paper
Interaction	Team BIWI	Yao et al.	ETH	△	Variations of a Hough-Voting Action Recognition System
	TU Graz	-	TU Graz	X	-
	SUVARI	-	Sabanci Univ.[1]	X	-
	Panopticon	-	Sabanci Univ.[1]	X	-
Aerial-view	Imagelab	Vezzani et al.	Univ. of Modena and Reggio Emilia	O	HMM based Action Recognition with Projection Histogram Features
	ECSI_ISI	Biswas et al.	Indian Statistical Institute	O	-
	BU_Action	Guo et al.	Boston University	O	Aerial View Activity Classification by Covariance Matching of Silhouette Tunnels
	Team BIWI	Yao et al.	ETH	O	Variations of a Hough-Voting Action Recognition System
Wide-area	Vistek	-	Sabanci Univ.[2], Univ. of Amsterdam	X	-

man Interaction Recognition Challenge, Aerial View Activity Classification Challenge, and Wide-Area Activity Search and Recognition Challenge. The three types of datasets named UT-Interaction, UT-Tower, and UCR-Videoweb are introduced for each challenge respectively. The objective of our challenges is to provide videos of human activities which are of practical interests, and make researchers evaluate their existing/new activity recognition methodologies with our real-world settings.

In the interaction challenge, contestants are asked to correctly localize ongoing activities from continuous video streams containing multiple human-human interactions (i.e. a high-level surveillance setting). The aerial view challenge requires the participants to develop recognition methodologies that handles low-resolution videos where each person's height is of approximately 20 pixels. This challenge is particularly motivated by military applications such as unmanned aerial vehicles taking videos from an aerial view. The wide-area challenge asks contestants to retrieve videos similar to query events using a multi-camera dataset. This dataset consists of videos obtained from multiple camera covering different regions of a wide area, which is a very common situation in many surveillance scenarios (e.g. airport).

The challenges are designed to encourage researchers to test their new state-of-the-art recognition systems on the three datasets with different characteristics (Table 2). Even though there exist other public datasets composed of human action videos [16, 8, 20, 13] (Fig. 3 (a-e)), most of them focus on recognition of simple actions (e.g. walking, jogging, ...) in controlled environments (e.g. only one actor appears in the videos, taken from a single camera). Several baseline methods have been implemented by the contest organizers as well, comparing contestants' results with well-known previous methodologies. The contest and its datasets will provide impetus for future research in many related areas.

Table 2. A table summarizing the characteristics of the contest datasets. '# Executions' describes the total number of activity executions in the entire dataset. '# Actors' is the number of actors appearing in the scene simultaneously, and 'Multi-person' describes whether the dataset involves multi-person activities or not.



Dataset Name	# Activities	# Executions	# Cameras	# Actors	Resolution	Multi-person	Continuous
UT-Interaction	6	120+	1	2~4	720*480	O	O
UT-Tower	9	108	1	1	360*240	X	X
UCR-Videoweb	52	Multiple	4~8	2~10	640*480	O	O

2 Previous Datasets

Several public datasets have been introduced in the past 10 years, encouraging researchers to explore various action recognition directions. The KTH dataset [16] and the Weizmann dataset [8] are the typical examples of these dataset. These two single-camera datasets have been designed for research purposes, providing a standard for researchers to compare their action classification performances. The datasets are composed of videos of relatively simple periodic actions, such as walking, jogging, and running. The videos are segmented temporally so that each clip contains no more than one action of a single person. They were taken in a controlled environment; their backgrounds and lighting conditions are mostly uniform. In general, they have a good image resolution and little camera jitters. The I-XMAS dataset [20] was similar, except that they provided videos from multiple cameras for a 3-D reconstruction.

Recently, more challenging datasets were constructed by collecting realistic videos from movies [13, 12, 14]. These movie scenes are taken from varying view points with complex backgrounds, in contrast of the previous public datasets [16, 8]. These dataset encourages the development of recognition systems that are reliable under noise and view point changes. However, even though these videos were taken in more realistic environments, the complexity of the actions themselves were similar to [16, 8]: the datasets contain simple instantaneous actions such as kissing and hitting. They were not designed to test recognition of high-level human activities from continuous sequences.

There also are datasets motivated by surveillance applications. PETS datasets [1] and i-LIDS datasets [6] belong to this category. The videos in these datasets were taken in uncontrolled environments (e.g. subway stations), and they contain few application specific activities (e.g. leaving a baggage). Videos from multiple cameras watching the same site with different view points are provided.

Each of the datasets introduced in SDHA 2010 has its unique characteristics that distinguish it from other previous datasets. The UT-Interaction dataset is designed to encourage detection of high-level human activities (e.g. hand shaking) which are more complex than previous simple actions. In addition, it encourages localization of the multiple activities from continuous video streams spatially and temporally. The UT-Tower dataset contains very low-resolution videos, which makes their recognition challenging. The UCR-Videoweb dataset introduces continuous videos taken from multiple cameras observing different areas of a place (e.g. CCTV cameras for a university campus building).

Up to our knowledge, SDHA 2010 is the first computer vision contest designed to compare performances of activity recognition methodologies. There have been previous competitions for recognizing objects (e.g. PASCAL-VOC challenges [7]) or recognizing a specific scene (e.g. abandoned baggage detection in AVSS 2007 [6]), but no previous contest attempted to measure general accuracies of systems on recognizing various types of human activities. SDHA 2010's objective is to test human activity recognition's state-of-the-arts and establish standard datasets for future exploration.

3 High-level Human Interaction Recognition Challenge

In the "High-level Human Interaction Recognition Challenge", contestants are asked to recognize ongoing human activities from continuous videos. The objective of the challenge is to encourage researchers to explore the recognition of complex human activities from continuous videos, taken in realistic settings. Each video contains several human-human interactions (e.g. hand shaking and pushing) occurring sequentially and/or concurrently. The contestants must correctly annotate which activity is occurring when and where for all videos. Irrelevant pedestrians are also present in some videos. Accurate detection and localization of human activities are required, instead of a brute force classification of videos.

The motivation is that many of real-world applications require high-level activities performed by multiple individuals to be recognized. Surveillance systems for airports and subway stations are typical examples. In these environments, continuous sequences provided from CCTV cameras must be analyzed toward correct detection of multi-human interactions such as two persons fighting. In contrast to previous single-person action classification datasets discussed in Section 2, the challenge aims to establish a new public dataset composed of continuous executions of multiple real-world human interactions.

3.1 Dataset Description

The UT-Interaction dataset[1] contains videos of continuous executions of 6 classes of human-human interactions: hand-shake, point, hug, push, kick and punch. Fig. 1 shows example snapshots of these multi-person activities. Ground truth labels for all interactions in the dataset videos are provided, including time intervals and bounding boxes. There is a total of 20 video sequences whose lengths are around 1 minute (e.g. Fig. 2). Each video contains at least one execution per interaction, providing us about 8 executions of human activities per video on average. Several actors with more than 15 different clothing conditions appear in the videos. The videos are taken with the resolution of 720*480, 30 fps, and the height of a person in the video is about 200 pixels.

We divide videos into two sets. The set #1 is composed of 10 video sequences taken on a parking lot. The videos of the set #1 are taken with slightly different

[1] http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html

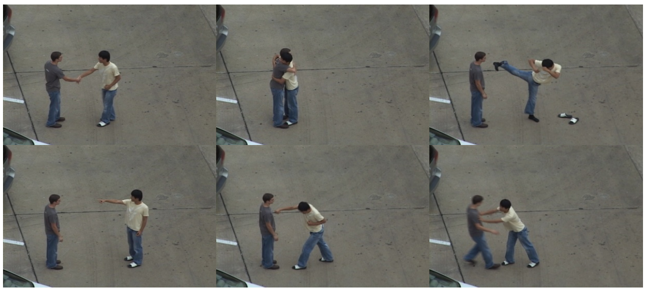


Fig. 1. Example snapshots of the six human-human interactions.

zoom rate, and their backgrounds are mostly static with little camera jitter. The set #2 (i.e. the other 10 sequences) are taken at a lawn on a windy day. Background is moving slightly (e.g. tree moves), and they contain more camera jitters. From sequences 1 to 4 and from 11 to 13, only two interacting persons appear in the scene. From sequences 5 to 8 and from 14 to 17, both interacting persons and pedestrians are present in the scene. In sets 9, 10, 18, 19, and 20, several pairs of interacting persons execute the activities simultaneously. Each set has a different background, scale, and illumination. The UT-Interaction set #1 was first introduced in [15], and we are extending it with this challenge.

For each set, we selected 60 activity executions that will be used for the training and testing in our challenge. The contestant performances are measured using the selected 60 activity executions. The other executions, marked as 'others' in our dataset, are not used for the evaluation.

3.2 Results

The interaction challenge consists of two types of tasks: the classification task and the continuous detection (i.e. localization) task. The contestants are requested to evaluate their systems with these two different experimental settings:

For the 'classification', 120 video segments (from 20 sequences) cropped based on the ground truth bounding boxes and time intervals are provided. The video sequences were segmented spatially and temporal to contain only one interaction performed by two participants, and the classification accuracies are measured with these video segments similar to previous settings [16, 8]. That is, the performance of classifying a testing video segment into its correct category is measured.

In the 'detection' setting, the entire continuous sequences are used for the continuous recognition. The activity recognition is measured to be correct if and only if the system correctly annotates an occurring activity's time interval (i.e. a pair of starting time and ending time) and its spatial bounding box. If the annotation overlaps with the ground truth more than 50% spatially and temporally, the detection is treated as a true positive. Otherwise, it is treated

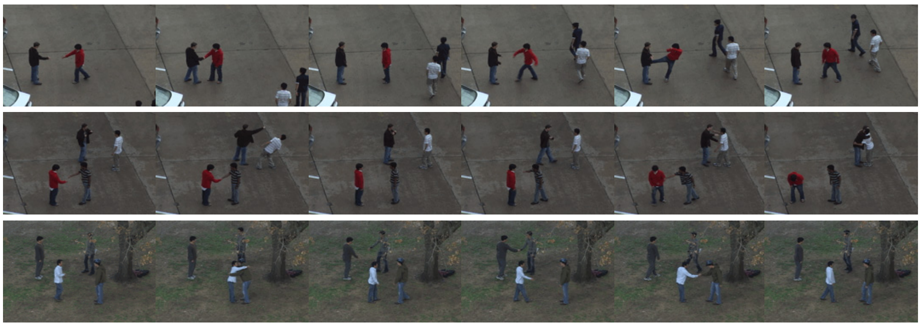


Fig. 2. Example video sequences of the UT-Interaction dataset.

as a false positive. Contestants are requested to submit a Precision-Recall curve for each set, summarizing the detection results.

In both tasks, the contestants were asked to measure the performances of their systems using 10-fold leave-one-out cross validation per set as follows: For each round, contestants leave one among 10 sequences for the testing and use the other 9 for the training. Contestants are required to count the number of true positives, false positives, and false negatives obtained through the entire 10 rounds, which will provide a particular precision and recall rate of the system (i.e. a point on a PR curve). Various PR rates will be obtained by changing the system parameters, and the PR curve is drawn by plotting them.

A total of four teams showed their intent to participate the challenge. However, only one among them succeeded to submit results for the classification task, which we report with Tables 3 and 4. The team BIWI [19] used a Hough transform-based method to classify interaction videos. Their method is based on [21], which uses a spatio-temporal voting with extracted local XYT features. A pedestrian detection algorithm was also adopted for the better classification. Particularly for the interaction challenge, they have modeled each actor's action using their voting method, forming a hierarchical system consisting of 2-levels.

In addition, in order to compare the participating team's result with previous methodologies, we have implemented several existing well-known action classification methods. Two different types of features (i.e. spatio-temporal features from [16] and 'cuboids' from [4]) are adopted, and three types of elementary classifiers, {k-nearest neighbor classifiers (k-NNs), Bayesian classifiers, and support vector machines (SVMs)}, are implemented. Their combinations generate six baseline methods as specified in Tables 3 and 4.

The baseline classifiers rely on a feature codebook generated by clustering the feature vectors into several categories. Codebooks were generated 10 times using k-means algorithm, and the systems' performances have been averaged for the 10 codebooks. SVM classification accuracies with the best codebook is also provided for the comparison. In the baseline methods, video segments have been normalized based on the ground truth so that the main actor of the activity (e.g. the person punching the other) always stands on the left-hand side.

Table 3. Activity classification accuracies of the systems tested on the UT-Interaction dataset #1. The 1st, 2nd, and 3rd best system accuracies are described per activity: the blue color is for the 1st, the orange color suggests the 2nd, and the green color is for the 3rd.

	Shake	Hug	Kick	Point	Punch	Push	Total
Laptev + kNN	0.18	0.49	0.57	0.88	**0.73**	0.57	0.57
Laptev + Bayes.	0.38	0.72	0.47	0.9	0.5	0.52	0.582
Laptev + SVM	0.49	0.79	0.58	0.8	0.6	0.59	0.642
Latpev + SVM (best)	0.5	0.8	0.7	0.8	0.6	0.7	0.683
Cuboid + kNN	0.56	0.85	0.33	0.93	0.39	0.72	0.63
Cuboid + Bayes.	0.49	0.86	**0.72**	**0.96**	0.44	0.53	0.667
Cuboid + SVM	**0.72**	**0.88**	**0.72**	0.92	0.56	**0.73**	**0.755**
Cuboid + SVM (best)	**0.8**	**0.9**	**0.9**	**1**	**0.7**	**0.8**	**0.85**
Team BIWI	**0.7**	**1**	**1**	**1**	**0.7**	**0.9**	**0.88**

Table 4. Activity classification accuracies of the systems tested on the UT-Interaction dataset #2.

	Shake	Hug	Kick	Point	Punch	Push	Total
Laptev + kNN	0.3	0.38	**0.76**	**0.98**	0.34	0.22	0.497
Laptev + Bayes.	0.36	0.67	0.62	0.9	0.32	**0.4**	0.545
Laptev + SVM	0.49	0.64	0.68	0.9	0.47	**0.4**	0.597
Latpev + SVM (best)	0.5	0.7	**0.8**	0.9	0.5	**0.5**	**0.65**
Cuboid + kNN	**0.65**	**0.75**	0.57	0.9	0.58	0.25	0.617
Cuboid + Bayes.	0.26	0.68	0.72	**0.94**	0.28	0.33	0.535
Cuboid + SVM	**0.61**	**0.75**	0.55	0.9	**0.59**	0.36	0.627
Cuboid + SVM (best)	**0.8**	**0.8**	0.6	0.9	**0.7**	**0.4**	**0.7**
Team BIWI	0.5	**0.9**	**1**	**1**	**0.8**	**0.4**	**0.77**

The classification results shows that the pointing interaction composed of least number of feature and the hugging interaction composed of the largest number of distinctive features was recognized with a high accuracy in general. Punching was confused with pushing in many systems because of their similarity. The participating team, BIWI, showed the highest recognition accuracy. The performances of the "Cuboid + SVM" with the best codebook were comparable.

3.3 Discussions

No team was able to submit a valid result for the detection task with continuous videos. There were 4 teams intended to participate challenge, but only one team succeeded to classify human interactions successfully and none succeeded to performed the continuous recognition. This implies that the recognition of high-level human activities from continuous videos still is an unsolved problem. Despite the demands from various applications including surveillance, the problem remains largely unexplored by researchers.

Applying the 'sliding windows' technique together with the classifier used above will be a straight forward solution. However, given the reported classification accuracies, such method is expected to generate many false positives. Using a voting-based methodology (e.g. [15, 21]) is a promising direction for the detection task, and they must be explored further. In addition, we were able to observe that the hierarchical approach obtained better performances than the other baseline methods. Developing hierarchical approaches for continuous detection and localization of complex human activities will be required.

4 Aerial View Activity Classification Challenge

The ability to accurately recognize human activities at a distance is essential for several applications. Such applications include automated surveillance, aerial or satellite video analysis, and sports video annotation and search, etc. However, due to perspective distortion and air turbulence, the input imagery is presented in low-resolution and the available action patterns tend to be missing and blurry. In addition, shadows, time-varying lighting conditions, and unstabilized videos can all add up to the difficulty of this task. Therefore, without explicitly addressing these issues, most existing work in activity recognition may not be appropriate under the scenario.

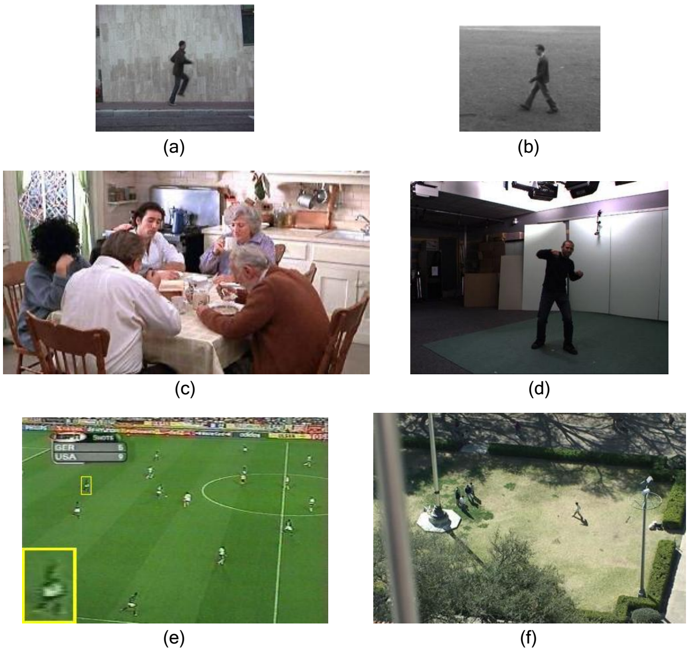
In this "Aerial View Activity Classification Challenge", we aim to motivate researchers to explore techniques that achieve accurate recognition of human activities in videos filmed from a distant view. To simulate the video settings, we took image sequences of a single person performing various activities from the top of the University of Texas at Austin's main tower. We name it UT-Tower dataset[2]. The average height of a human figure in this dataset is about 20 pixels. The contest participants are expected to classify 108 video clips from a total of 9 categories of human activities. The performance of each participating team is evaluated by their leave-one-out accuracy on the dataset.

As described in Section 2, there exist several public datasets that are widely referred and tested in the literature of human activity recognition [8, 16, 13, 5, 20]. However, all these datasets (except the Soccer dataset) are taken from an approximate side view and they have human figures presented in high-resolution imagery (Fig. 3). The Soccer dataset contains low-resolution videos similar to ours, but the action categories of the Soccer dataset are defined by the proceeding directions of the players, and nearly half of the video sequences are the mirrors of the other half. These issues limit their applicability to the evaluation of activity recognition algorithms that focus on low-resolution video settings. Therefore, with this challenge, we distribute a new dataset for the assessment of general and surveillance oriented applications.

4.1 Dataset Description

Filmed top-down from a 307-foot high tower building, the UT-Tower dataset is composed of low-resolution videos similar to the imagery taken from an aerial vehicle. There are 9 classes of human actions: 'pointing', 'standing', 'digging', 'walking', 'carrying', 'running', 'wave1', 'wave2', 'jumping'. Algorithm performance on both *still* and *moving* types of human activities are to be examined.

[2] http://cvrc.ece.utexas.edu/SDHA2010/Aerial_View_Activity.html



(a) (b) (c) (d) (e) (f)

Fig. 3. The widely used public datasets are mostly in medium- to high-resolution, for example, (a) Weizmann dataset, (b) KTH dataset, (c) HOHA dataset, and (d) I-XMAS dataset. Low-resolution datasets include (e) Soccer dataset and the proposed (f) UT-Tower dataset. The sizes of the images are proportional to their actual resolutions.

A total of 6 individuals acted in this dataset. We let each performer repeat every activity twice so that there are 108 sequences in the dataset. To add to the variety of the dataset, we recorded the activities under two types scene settings: concrete square and lawn. The videos were taken in 360×240 pixels resolution at 10fps. In addition to the low-resolution setup, the UT-Tower dataset also poses other challenges. For example, the direct sunlight causes salient human cast shadows and the rooftop gust brings continuous jitters to the camera. Fig. 4 shows the example video sequences of the dataset.

We manually segmented the original video into short clips so that each clip contains one complete track of human activity. In order to alleviate segmentation and tracking issues and make participants focus on the classification problem, we provide ground truth bounding boxes as well as foreground masks for each video. Contestants are free to take advantages of them or apply their own preprocessing techniques.

4.2 Results

In the aerial challenge, the contestants were asked to classify video clips in the UT-Tower dataset into the above-mentioned 9 action categories. Similar to the

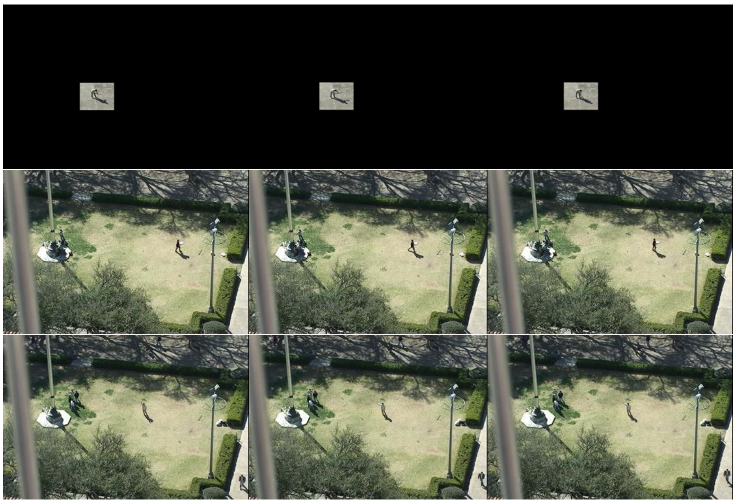


Fig. 4. The examples of 'digging', 'carrying', and 'wave1' in the UT-Tower dataset.

classification task of the interaction challenge, a leave-one-out cross validation setting is used. Here, one among 108 videos are used for the testing, and the others are used for the training. That is, a 108-fold cross validation is performed to evaluate the performances of the systems.

There are totally 4 university teams participated in this contest. Each team has tested their proposed algorithm on the UT-Tower dataset and reported the results. We briefly summarize the submitted methodologies and our baseline technique as follows.

Team BIWI BIWI team from ETH Zurich proposes to use a Hough transform-based voting framework for action recognition [19]. They separate the voting into two stages to bypass the inherent high dimensionality problem in Hough transform representation. Random trees are trained to learn a mapping between densely-sampled feature patches and their corresponding votes in a spatio-temporal-action Hough space. They perform recognition by voting with a collection of learned random trees.

BU Action Covariance Manifolds Boston University team represents a track of human action as a temporal sequence of local shape-deformations of centroid-centered object silhouettes [10], i.e., the shape of the silhouette tunnel. The empirical covariance matrix of a set of 13-dimensional feature is extracted as feature from the silhouette tunnel. The silhouette tunnel of a test video is broken into short overlapping segments and each segment is classified using a dictionary of labeled action covariance matrices with the nearest neighbor rule.

ECSU_ISI The team from Indian Statistical Institute adopts a bag-of-word-based approach, which represents actions by the chosen key poses. The key

Table 5. System accuracies (%) of the aerial-view challenge.

	Point	Stand	Dig	Walk	Carry	Run	Wave1	Wave2	Jump	Total
Team BIWI	100	91.7	100	100	100	100	83.3	83.3	100	95.4
BU	91.7	83.3	100	100	100	100	100	100	100	97.2
ECSU_ISI	100	83.3	91.7	100	100	100	100	91.7	91.7	95.4
Imagelab	83.3	83.3	100	100	100	100	100	100	100	96.3
Baseline	100	83.3	100	100	100	100	83.3	100	100	96.3

poses are extracted from an over-complete codebook of poses using the theory of graph connectivity. They train a Support Vector Machines (SVM) classifier to perform action classification.

Imagelab The University of Modena and Reggio Emilia team applies a hidden Markov model (HMM) based technique [18] on the dataset. Their action descriptor is a K-dimensional feature set extracted from the projection histograms of the foreground masks. They train a HMM per action and is able to perform recognition on-line.

Baseline We consider a baseline approach as a simple combination of a commonly used feature and a linear classifier. For this purpose, we use time serious of Histogram of Oriented Gradients (HOG) [3] to characterize successive human poses and a linear kernel SVM classifier for classification. A track of human action is divided into overlapped spatio-temporal volumes, from which we extract and concatenate sequences of HOG vectors as the baseline action descriptors.

We tabulate the average accuracy per activity as well as the overall accuracy of each team and the baseline method in Table 5. Note that prior to this competition, Chen and Aggarwal [2] have tested their method on part of this dataset (60 sequences of the lawn scene). They were able to achieve 100% accuracy on the partial dataset. For the sake of fairness, we did not include their latest results in this paper.

4.3 Discussions

As shown in Table 5, all the contestants achieve very similar accuracies on this low-resolution dataset. The BU team using a silhouette-based method performed the best among four participating teams. In addition, we are surprised to find out that the baseline method was comparable; it obtained the 2nd best performance. 'pointing', 'standing', 'wave1', and 'wave2' are the most common activities that caused misclassifications. The action pairs of <pointing, standing>, <pointing, wave1>, and <wave1, wave2> can be confusing to a recognition algorithm in the sense that one action can only be distinguished from the other by a very short period of hand motion. In low-resolution imagery, vague and sparse action features, salient human cast shadow, and unstabilized videos can all make the discerning task even more challenging. Therefore, we believe a more elaborate preprocessing procedure and the employment of multiple features in classification may further the performance on this dataset.

5 Wide-Area Activity Search and Recognition Challenge

The objective of the "Wide-Area Activity Search and Recognition Challenge" is to search a video given a short query clip in a wide-area surveillance scenario. Our intention is to encourage the development of activity recognition strategies that are able to incorporate information from a network of cameras covering a wide-area. The UCR-Videoweb dataset[3] introduced in this paper has activities that are viewed from 4-8 cameras and allows us to test performance in a camera network. For each query, a clip video containing a specific activity was provided, and the contestants are asked to search for similar videos.

In contrast to the other two challenges, the wide-area challenge was an open challenge: The contestants were free to choose particular types of human activities from the dataset for the recognition, and they were allowed to explore a subset of entire videos.

5.1 Dataset Description

The Videoweb dataset consists of about 2.5 hours of video observed from 4-8 cameras. The data is divided into a number of scenes that were collected over several days. Each scene is observed by a camera network where the actual number of cameras changes depending on the scene due to its nature. For each scene, the videos from the cameras are available. Annotation is available for each scene and the annotation convention is described in the dataset. It identifies the frame numbers and camera ID for each activity that is annotated. The videos from the cameras are approximately synchronized.

The videos contain several types of activities including throwing a ball, shaking hands, standing in a line, handing out forms, running, limping, getting into/out of a car, and cars making turns. The number for each activity varies widely. The data was collected in 4 days and the number of scenes are: {day1: 7 scenes}, {day2: 8 scenes}, {day3: 18 scenes}, and {day4: 6 scenes}. Each scene are on average 4 minutes long and there are 4-7 cameras in each scene. Each scene contains multiple activities. Figure 6 shows example sequences of the dataset.

5.2 Results

In the wide-area challenge, the contestants were asked to formulate their own activity search problem with the dataset, and report their results. That is, each contestant must choose query clips from some scenes in the dataset and use them to retrieve similar scenes in another parts of the dataset. The 'correctly identified clip' is defined as the clip in which the overlap in the range of frame numbers obtained by the search engine for an activity is at least 50% of the range in the annotation and not more than 150% of that range.

There was a single team who showed an intention to participate the wide-area challenge. However, unfortunately, no team succeeded to submit valid results for

[3] http://vwdata.ee.ucr.edu/

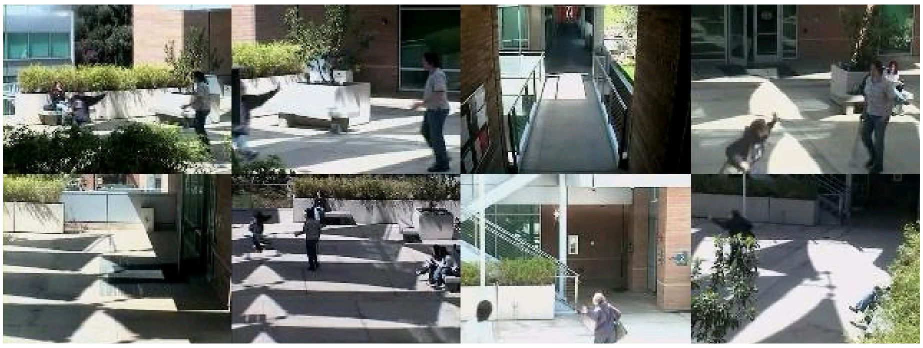


Fig. 5. Example images from the UCR-Videoweb dataset. Each image shows a snapshot obtained from one of 8 different cameras at a particular time frame.

the wide-area challenge. Here, we report results of systems implemented by the contest organizer [11], so that they can be served as a baseline for the future research. We formulate three types of problems, where each of them focuses on the search of different types of human activities, and report the system performances on these tasks.

Query-based Complex Activity Search In this task, we searched for interactions in videos using a single video clip as a query. We worked with 15 minutes of video where up to 10 different actors take part in any given complex activity which involves interaction of humans with other humans, objects, or vehicles. We have used 6 scenes from day 3 data as the test set. The problem was very similar to the human-human interaction detection problem in Section 3, recognizing three types of interactions: shaking hands, hugging, and pointing. Table 6 shows the detection accuracies together with false positive rates.

Table 6. Recognition accuracy on three complex human-human interactions.

Interaction	**Our recognition accuracy**	**False positive rate**
Shake hands	0.68	0.57
Hug	0.74	0.55
Point	0.63	0.25

Modeling and Recognition of Complex Multi-Person Interactions in Video This task is to examine the formation and dispersal of groups and crowds from multiple interacting objects, which is a fast-growing area in video search. We search for activities involving multiple objects and analyze group formations and interactions. For this task, four scenes have been used for the testing (more details can be found at [11]). We apply a modeling-based methodology to test the implemented system within a query-based retrieval framework. Table 7 shows the types of interactions searched and the precision/recall values of the system.

Table 7. Precision/Recall Values for DB query and retrieval of two-object and complex, multi-object motions.

Activity	**Precision**	**Recall**	**Total Fetched**	**True Pos.**	**Ground Truth**
Person Entering Building	1	1	4	4	4
Person Exiting Building	1	1	2	2	2
Person Entering Vehicle	0.75	0.75	4	3	3
Person Exiting Vehicle	1	1	3	3	3
People Walking Together	1	0.6	3	3	5
People Coming Together	0.7	0.7	7	5	5
People Going Apart	0.8	1	5	4	5
People Milling Together	0.78	0.92	14	11	13
People Meandering Together	0.85	0.92	27	23	25
Group Formation	1	0.78	7	7	9
Group Dispersal	0.8	0.8	5	4	4
Person Joining Group	1	0.95	18	18	19
Person Leaving Group	1	1	11	11	11

Activity Recognition Based on Multi-camera Data-fusion In the past few years, multi-camera installations have rapidly positioned themselves in many applications, e.g., video surveillance, national and homeland security, assisted living facilities, environmental monitoring, disaster response etc. The automated analysis of human actions from these video streams has gained a lot of importance recently. The goal of this task is to search for human actions in such an environment, integrating information from multiple cameras.

We used 8 scenes from day3: the segments of videos having at least one of our defined action classes were selected from these 8 scenes. 10 minutes of the UCR-Videoweb data-set was used for training and another 10 minutes was used for testing. We trained our system for six different action classes, i.e. 1 - Sit, 2 - Walk, 3 - Picking up object, 4 - Shake hand, 5 - Hug and 6 - Wave one hand. Approximately 15 video clips of 2-3 seconds each were used to train our classifier per activity. For each action class, 10-20 instances of each action were used for testing and about 30 different scenarios of multiple actions occurring in multiple cameras were used for testing.

We show the statistics of the performance gain of our method over single-view action recognition scores in Fig. 6. That is, we show that data association and information fusion among multiple cameras improves recognition performance.

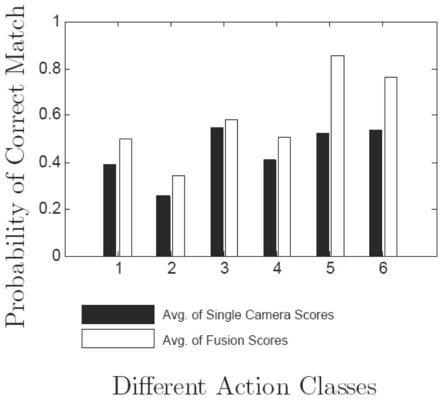


Fig. 6. This figure shows the comparison of the recognition scores of our overall approach with single camera action recognition scores. For action class 3, the single view action recognition was almost flat over all action classes, so the fusion could not improve the result much. On the other hand, in action class 5, at least one of the cameras got a good shot of the action and the fused scores went up. In this experiment, each of the targets was viewed by 1-3 cameras simultaneously.

6 Conclusion

In this overview paper, we have summarized the results of the first Contest on Semantic Description of Human Activities (SDHA) 2010. SDHA 2010 is one of the very first activity recognition contest, consists of three types of challenges. The challenges introduced three new public datasets (UT-Interaction, UT-Tower, and UCR-Videoweb), which motivated contestants to develop new approaches for complex human activity recognition scenarios in realistic environments. Researchers from various universities participated in SDHA 2010, proposing new activity recognition systems and discussing their results. In addition, several baseline methods were implemented and compared with contestants' results. SDHA 2010 evaluated the human activity recognition's state-of-the-arts.

Table 1 summarizes the results of SDHA 2010. A total of four teams showed their intent to participate the interaction challenge. However, only a single team succeeded to submit results for the classification task, and no team submitted correct detection results. There were four teams participated in the aerial-view challenge, and all teams submitted results with high recognition accuracies (>0.95). One team intended to participated the wide-area challenge, but the team decided not to submit the results. This is due to the fact that the activities used in the aerial-view challenge were relatively simple compared to the others. Simple one-person actions were classified in the challenge, while the activities in the other two challenges include high-level multi-person interactions. We are able to observe that localization of ongoing activities from continuous video streams is a challenging problem, which remains open for future investigations.

References

1. PETS 2006 benchmark data. http://www.cvg.rdg.ac.uk/PETS2006/data.html
2. Chen, C.C., Aggarwal, J.K.: Recognizing human action from a far field of view. In: IEEE Workshop on Motion and Video Computing (WMVC) (2009)
3. Dalal, N., Triggs., B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
4. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: IEEE VS-PETS Workshop. pp. 65–72 (2005)
5. Efros, A.A., Berg, A.C., Mori, G., Malik., J.: Recognizing action at a distance. In: ICCV (2003)
6. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: i-LIDS abandoned baggage detection challenge dataset. http://www.elec.qmul.ac.uk/staffinfo/andrea/avss2007_ss_challenge.html
7. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results. http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html
8. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. PAMI 29(12), 2247–2253 (December 2007)
9. Guo, K., Ishwar, P., Konrad, J.: Action change detection in video by covariance matching of silhouette tunnels. In: ICASSP (2010)
10. Guo, K., Ishwar, P., Konrad, J.: Action recognition in video by sparse representation on covariance manifolds of silhouette tunnels. In: ICPR contest on Semantic Description of Human Activities (SDHA), in Proceedings of the ICPR contests (2010)
11. Kamal, A., Sethi, R., Song, B., Fong, A., Roy-Chowdhury, A.: Activity recognition results on UCR Videoweb dataset. In: Technical Report, Video Computing Group, University of California, Riverside (2010)
12. Ke, Y., Sukthankar, R., Hebert, M.: Spatio-temporal shape and flow correlation for action recognition. In: CVPR (2007)
13. Laptev, I., Perez, P.: Retrieving actions in movies. In: ICCV (2007)
14. Rodriguez, M.D., Ahmed, J., Shah, M.: Action MACH: A spatio-temporal maximum average correlation height filter for action recognition. In: CVPR (2008)
15. Ryoo, M.S., Aggarwal, J.K.: Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: ICCV (2009)
16. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local SVM approach. In: ICPR (2004)
17. Vezzani, R., Piccardi, M., Cucchiara, R.: An efficient bayesian framework for online action recognition. In: ICIP (2009)
18. Vezzani, R., Baltieri, D., Cucchiara, R.: HMM based action recognition with projection histogram features. In: ICPR contest on Semantic Description of Human Activities (SDHA), in Proceedings of the ICPR contests (2010)
19. Waltisberg, D., Yao, A., Gall, J., Gool, L.V.: Variations of a Hough-voting action recognition system. In: ICPR contest on Semantic Description of Human Activities (SDHA), in Proceedings of the ICPR contests (2010)
20. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. CVIU 104(2), 249–257 (2006)
21. Yao, A., Gall, J., Gool, L.V.: A hough transform-based voting framework for action recognition. In: CVPR (2010)