The i3DPost multi-view and 3D human action/interaction database

Nikolaos Gkalelis[2 3], Hansung Kim[1], Adrian Hilton[1], Nikos Nikolaidis[2 3] and Ioannis Pitas[2 3]

[1] University of Surrey, Guildford, UK
[2] Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece
[3] Department of Informatics, Aristotle University of Thessaloniki, Greece

Abstract

In this paper a new multi-view/3D human action/interaction database is presented. The database has been created using a convergent eight camera setup to produce high definition multi-view videos, where each video depicts one of eight persons performing one of twelve different human motions. Various types of motions have been recorded, i.e., scenes where one person performs a specific movement, scenes where a person executes different movements in a succession and scenes where two persons interact with each other. Moreover, the subjects have different body sizes, clothing and are of different sex, nationalities, etc.. The multi-view videos have been further processed to produce a 3D mesh at each frame describing the respective 3D human body surface. To increase the applicability of the database, for each person a multi-view video depicting the person performing sequentially the six basic facial expressions separated by the neutral expression has also been recorded. The database is freely available for research purposes.

1 Introduction

The vast majority of the human action recognition methods use single-view video sources and pose the requirement that the human is captured from the same viewing angle during both the testing and training stage, e.g., all walking videos should depict the lateral side of the human. Experimental results have shown that these algorithms can tolerate only a small deviation from the training view angle, and are susceptible to partial occlusion. Consequently, it is expected that full view invariant action recognition, robust to occlusion, will be much more feasible through algorithms based on multi-view videos or 3D posture model sequences.

Multiple view video and 3D technology is currently attracting growing attention in several disciplines, for instance, in the entertainment industry [8], where it can be used to provide high quality multi-perspective viewing experiences and 3D scene/actor reconstructions for digital cinema movies and interactive games, in security applications [20], for view independent non-invasive event detection, and in other areas. Human action recognition could play an important role in such applications, e.g., by providing "anthropocentric" semantic information for the characterization, summarization, indexing and retrieval of multi-view and 3D data, or for the detection of unusual activities in video surveillance systems.

Several taxonomies have been proposed to aid the analysis of human motion [1, 3, 4, 11]. Inspired from these taxonomies we propose the taxonomy shown in Figure 1. In the lowest level, a dyneme is the most elementary constructive unit of motion, while one level above, the movement is a unique sequence of dynemes with some contextual meaning, e.g., a walking step. In the corresponding literature, human activity recognition and human action recognition have been often used interchangeably, i.e., to declare an algorithm that recognizes a sequence of movements, e.g, several steps of walk, run, etc.. For this reason, at the next level of the hierarchy, we use both the above forms to define the type of human motion that consists of a sequence of movements. The activity/action of walking that consists of several walking steps is such an example. In the same level of the hierarchy we also place human interactions despite being usually more complex than actions. Within this set, we may have interaction of a person with his environment, interaction between two or more people, or, in the most complex scenario, interaction of a group of people with each other as well as with their environment. Within interactions, people may perform single movements or sequence of movements, which may constitute specific activities. To depict the latter relation, a line is drawn from the set of interactions to the set of activities.

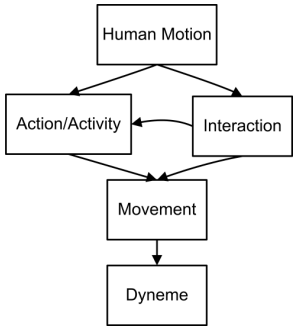


Figure 1: Human motion taxonomy.

The most recent survey regarding human action recognition

978-0-7695-3893-8/09 $26.00 © 2009 IEEE
DOI 10.1109/CVMP.2009.19

using video sources is reported in [23]. Here we provide a short review of existing multi-view video databases as well as of the most popular single-view databases used for the training and evaluation of human action recognition algorithms. In what concerns single-view databases, two publicly available ones, the Weizmann [10] and the KTH [19] database, have been widely used for the evaluation and comparison of single-view video human action recognition algorithms. The Weizmann database contains low resolution videos of nine persons performing ten actions, namely, "walk", "run", "skip", "gallop sideways", "jump jack", "jump forward", "jump in place", "bend", "wave with one hand" and "wave with two hands". Additionally, it contains twenty "robustness" videos, depicting a person walking under various scenarios, e.g., behind an obstacle, carrying a bag, etc.. The KTH database includes 25 persons and six actions: "walking", "jogging", "running", "boxing", "hand waving" and "hand clapping". Each person executed the same action several times under four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. Thus, in total the database contains 2391 sequences.

With respect to multi-view and 3D activity databases, many researches have constructed their own databases for verifying their algorithms, eg., [12, 13, 16, 24, 27]. In [12, 13] two orthogonally placed cameras are used for the creation of a two-view database of seven actors and nine movements per actor. Two cameras are used as well in [25] to record the frontal and lateral view for six different gymnastic activities and fifty persons. In [16] a convergent eight camera setup is used to create multi-view training videos for ten subjects and ten actions as well as one interaction. In [27] six cameras are used to record two subjects (male/female) performing eleven different actions. The silhouettes of the subjects from the six-view videos are extracted and used to to compute the visual hull of the human at each video frame and thus provide the respective rough 3D human body models. All the above described databases are not available to the public.

A database that is available for a fee has been reported in [7] and used to verify the algorithms proposed in [2, 14]. This database, named full body gesture (FBG) database, contains a total of 24 normal/abnormal actions and 20 persons. For each recording, three types of data are provided, i.e., stereo video data generated using 3 sets of stereo cameras, 3D motion capture data in the form of joint angle values over time for a skeleton model, and 2D human body binary mask sequences created using the stereo cameras. The INRIA Xmas Motion Acquisition Sequences (IXMAS) database, reported in [28], is probably the only publicly available multi-view/3D database. This database contains five-view video and 3D body model sequences for eleven actions and ten persons. It has been used for benchmarking the algorithms in [15, 26, 29]. Another popular database used for the verification of 3D human action recognition algorithms is the CMU database [6]. Although publicly available, this database contains only motion capture data (joint angles over time for a skeleton) a fact that restricts its usage [17, 18]. However, efforts have been made to generate artificial multi-view video data and 3D body models from the motion capture data, e.g. as done in [9]. Table 1 provides a concise summary for the reported 3D and multi-view databases.

The limited public availability of multi-view or 3D activity databases is definitely a factor that hampers the development of multi-view algorithms in this area. In this paper we describe a new publicly available database for the training, verification and evaluation of multi-view/3D human action recognition algorithms, hoping to aid in the development of this research topic. This database contains eight-view videos as well as their associated 3D posture model sequences (in the form of triangle meshes) for twelve natural actions and interactions. Binary body masks are also provided for all frames and views. The database has been developed within the European Union R&D project i3DPost and thus bears its name. To the best of our knowledge, this is the only multi-view/3D database that provides interaction videos as well as videos describing actions that contain more than one movements in succession.

2 Multi-view and 3D human action/interaction database

In this section we describe the camera setup the database content and the processing steps followed in order to derive the respective binary body masks and the 3D posture models.

2.1 Studio environment and camera set up

The studio where the database was recorded is equipped with eight Thomson Viper cameras, equally spaced in a ring of 8m diameter at a height of 2m above the studio floor. An even ambient illumination of around 4000 lux is provided by an array of KinoFlo fluorescent tubes on the ceiling, with flickerless operation and a consistent color spectrum. The cameras were positioned above the capture volume and were directed downward to exclude the lighting from the field-of-view. The cameras have a wide 45^{o} baseline to provide 360^{o} coverage of a capture volume of $4m \times 4m \times 2m$. A blue-screen backdrop was used to facilitate foreground segmentation (Figure 2). Human actions are captured in HD-SDI 20-bit 4:2:2 format with 1920×1080 resolution at 25Hz progressive scan. Synchronized videos from all eight cameras are recorded uncompressed direct to disk with eight dedicated PC capture boxes using DVS HD capture cards. Figure 3 shows an example of the captured multi-view videos.

All cameras were calibrated to extract their intrinsic (focal length, centre-of-projection, radial distortion) and extrinsic (pose, orientation) parameters before shooting. To achieve rapid and flexible multiple camera calibration we use a wand-based calibration technique [21]. This technique allows calibration of studio camera systems in less than 10 minutes with an accuracy comparable to normal grid-based calibration.

Camera calibration data are provided in an ASCII text file. This file defines the following camera parameters:

{number of cameras = 8} {distortion model = 1}

Database	Type	Contained action/interaction	# Motions	# Persons	# Cameras
[12]	private	Point to, raise hand, wave hand, touch head, communication, bow, pick up, kick, walk.	9	8	6
[25]	private	Gymnastic activities described in detail in the paper.	6	50	2
[16]	private	Walk, jump, pick up, kick, kneel, squat, punch, turn, sit, wave, handshake.	11	10	8
[27]	private	Lift right arm ahead, lift right arm sideways, lift left arm sideways and ahead, lift left arm sideways, lift both arms ahead then sideways, drop both arms sideways, lift both arms sideways, lift right leg bend knee, lift left leg bend knee, lift right leg firm, jump.	11	2	6
FBG [7]	commercial	Normal actions: sitting on a chair, standing up from a chair, walking in place, bending torso at waist, raising the right arm, extending hand for handshake, bowing from waist, sitting on the floor, getting down on the floor, lying down on the floor, waving the hand, running in place, walking forward, walking circularly. Abnormal actions: falling forward (standing), falling backward (standing), falling leftward (standing), falling rightward (standing), falling leftward (sitting on the floor), falling rightward (sitting on the floor), falling backward (sitting on the floor), falling leftward (sitting on a chair), falling rightward (sitting on a chair), falling forward (sitting on a chair).	24	20	3 (stereo pairs)
IXMAS [28]	public	Check-watch, cross-arms, scratch-head, sit-down, get-up, turn-around, walk, wave, punch, kick, pick-up.	11	10	5
CMU [6]	public	Motion capture data for 2605 trials in 6 categories and 23 subcategories. The main categories are: human interaction, interaction with environment, locomotion, physical activities & sports, situations & scenarios, test motions.			mocap data
i3DPost	public	Walk, run, jump forward, jump in place, bend, one hand wave, sit down - stand up, walk - sit down, run - fall, run - jump - walk, two persons handshaking, one person pulls another.	12	8	8

Table 1: *Multi-view/3D human motion databases.*

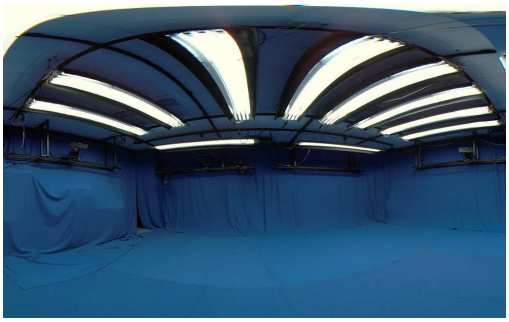


Figure 2: 3D Production Studio.

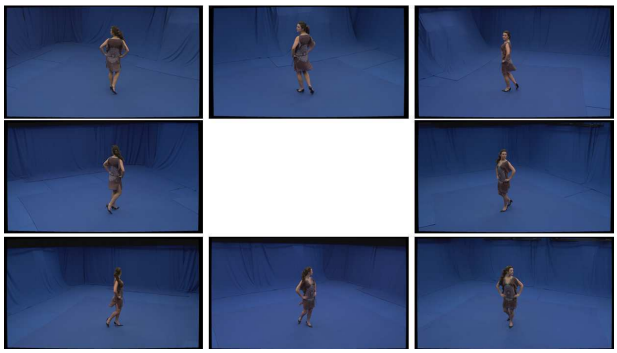


Figure 3: A single frame from the 8 videos.

then for each camera 00, 01, 02, ... 07

{image min row} {max row} {min column} {max column}
{focal length f_x} {f_y} {centre of projection c_x} {c_y}
{distortion k_1}
{r_{00}} {r_{01}} {r_{02}}
{r_{10}} {r_{11}} {r_{12}}
{r_{20}} {r_{21}} {r_{22}}
{t_0} {t_1} {t_2}
...

where in the above format

$$R = \begin{bmatrix} r_{00} & r_{01} & r_{02} \\ r_{10} & r_{11} & r_{12} \\ r_{20} & r_{21} & r_{22} \end{bmatrix} \quad \text{and} \quad t = \begin{bmatrix} t_0 \\ t_1 \\ t_2 \end{bmatrix}$$

are the camera rotation matrix and translation vector respectively.

The world coordinate frame for the cameras is defined on the floor at the centre of the capture volume. The coordinate system uses Y as the vertical axis.

The image plane of a camera has its origin at top left with coordinates (u, v) corresponding to the (column,row) of an image pixel. The projection (u, v) for a point $\mathbf{x}$ in world coordinates is defined as follows:

$$\begin{aligned} x_c &= Rx + t\,, \\ d_x &= f_x x_c[0]/x_c[2]\,, \\ d_y &= f_y x_c[1]/x_c[2]\,, \\ r &= \sqrt{(d_x^2 + d_y^2)}\,, \\ u &= c_x + d_x(1 + k_1 r)\,, \\ v &= c_y + d_y(1 + k_1 r)\,. \end{aligned}$$

2.2 Database description

Using the camera setup described above, eight amateurs (2 females and 6 males), participated in 13 distinct recording sessions each. Consequently, the database contains 104 multi-view videos or 832 (8×104) single view videos. In the following paragraphs we briefly describe each action or interaction. Next to the name of each action a two letters label is provided.

2.2.1 Activities containing instances of one specific movement

This category contains single movement activities, i.e., activities where a person performs sequentially several instances of the same movement, e.g., several walking steps. The activities of this type that are contained in the database are "walk" (wk), "run" (rn), "jump forward" (jf), "jump in place" (jp), "bend" (bd) and "one hand wave" (hw). Regarding "walk", "run" and "jump forward", the person is instructed to execute the activity from one side of the studio to the other, through the center of the capture volume. For both "bend" and "one hand wave" the person is in neutral position (i.e., "standing still") in the center of the capture volume. For the former, the person bends to pick up from the ground a fictional object and returns to the initial position, while for the latter the person is waving to a fictional person located at the side of the studio. Several frames for each action are depicted in Figure 4.

2.2.2 Activities containing instances of different movements

For the investigation of more complex scenarios as well as the study of human body dynamics involved during the transition from one movement type to another, four activity videos that contain instances of different movement types have been recorded for each subject. These activities are the following: "sit down - stand up" (ss), "walk - sit down" (ws), "run - fall" (rf) and "run - jump - walk" (rw). In the first activity scenario, "sit down - stand up", the person is initially in neutral position ("stand still"), positioned at the center of the capture volume and slowly bends his/her knees in order to take the "sit down" position. He/she stays in this position for a few seconds and then slowly returns back to the neutral position. In the other activities the subject is placed in neutral position in some distance outside the capture volume and then starts moving heading towards the center of the volume. Regarding the activity "walk - sit down" the person passes through the center of the capture volume and before reaching the border of the volume executes the "sit down" movement. In the "run - fall" activity, short after entering the capture volume, the subject is supposed to encounter a fictional obstacle and acts like losing his/her balance and consequently falling down. Finally, concerning the activity "run - jump - walk" when the person reaches the center of the capture volume executes a few instances of the "jump in place" movement and then continues walking in the same direction towards the edges of the studio. A few frames of each action are shown in Figure 5.

2.2.3 Interactions

Recognition of human interactions using multi-view or even single-view videos is a relatively unexplored topic. To aid research efforts in this direction, two multi-view interaction videos have been recorded for each person in the database. The first type of interaction included in the database is "two persons handshaking" (hs). In this scenario, two people are outside the capture volume and start walking towards each other. When they arrive in reach distance, they raise their hands and start handshaking for a short period of time. In the second interaction scenario, "one person pulls another" (pl), the two subjects are now located close to each other, near the center of the capture volume. During recording, one of the persons grabs the hand of the other and pulls him/her towards his/her side. A few frames extracted from interaction videos are shown in Figure 6.

2.2.4 Facial expression sequences

Before the recording session, each person watches a video where a trained actor is performing the six basic facial expressions. Afterwards, the subject is positioned at the edge of the capture volume, facing a specific camera, and he/she is instructed to sequentially execute the six facial expressions separated by the neutral expression, i.e., the person performs the following sequence of expressions: neutral - anger - neutral - disgust - neutral - fear - neutral - happiness - neutral - sadness - neutral - surprise. During recording, all or at least a large portion of the facial surface is captured from the five neighboring cameras in front of the subject. Due to the high resolution of the recorded videos the number of pixels on the facial area is sufficient high (approximately 3000 pixels). Figure 7 depicts a few frames extracted from a facial expression multi-view video, where a subject captured from five adjacent cameras poses the expression of happiness, while the cropped facial regions for the frames produced by the camera that captures the frontal facial view are depicted in Figure 8.

2.3 Multi-view video preprocessing

The captured videos are provided either as a set of images (one per frame) or in uncompressed avi format and are archived with the following naming convention: "ppp_aa_ccc.avi", where "ppp" declares the name of the person in the video, aa denotes the performed action in the video (in the two letter format described in the previous sections) and "ccc" is the index of the camera in the camera setup, e.g., "nik_bd_001.avi" denotes that the video was captured by camera one and depicts "nik" performing the movement bend. The raw video data were preprocessed to further increase the applicability of the database. Background substraction was applied by

Figure 4: *From top to bottom row, five frames of the following actions are shown: "walk" (wk), "run" (rn), "jump forward" (jf), "jump in place" (jp), "bend" (bd), "one hand wave" (hw).*

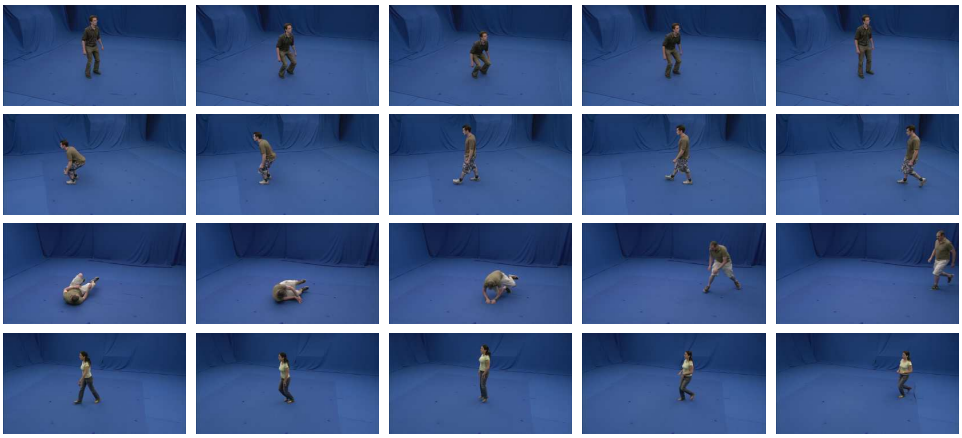


Figure 5: *From top to bottom row, five frames of the following actions are shown: "sit down - stand up" (ss), "walk - sit down" (ws), "run - fall" (rf) and "run - jump - walk" (rw).*

thresholding on the blue channel in order to produce binary mask sequences. These videos were archived with "_msk" appended to their name, e.g., "nik_bd_001_msk.avi". Each binary mask frame was further preprocessed to produce body posture regions of interest (ROIs) which contain as much foreground as possible. Due to the different size of the ROIs, a Matlab structure for each video was created and stored as mat file. The extension "_roi" was added to the name of the videos in this case, e.g., "nik_bd_001_roi.mat".

2.4 3D posture model reconstruction

The 3D posture model reconstruction is divided into several steps. First, the silhouettes are used to derive the visual-hull. The visual-hull defines an upper-bound on the true volume of the scene and so constrains the feasible space for surface reconstruction.

Shape-from-silhouette (SFS) in Fig 9 is a popular technique for scene reconstruction due to the simplicity of the reconstruction algorithm and the robust nature of shape recovery in the

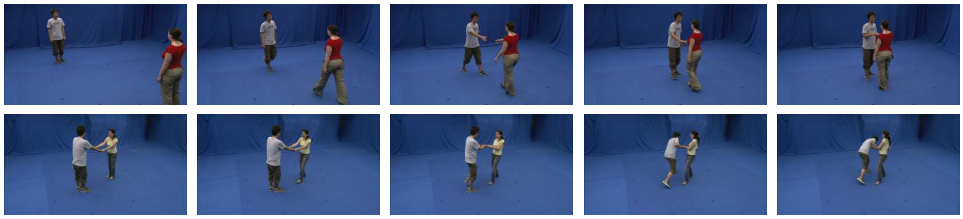


Figure 6: *From top to bottom row, five frames of the following actions are shown: "two persons handshake" (hs), "one person pulls another" (pl).*

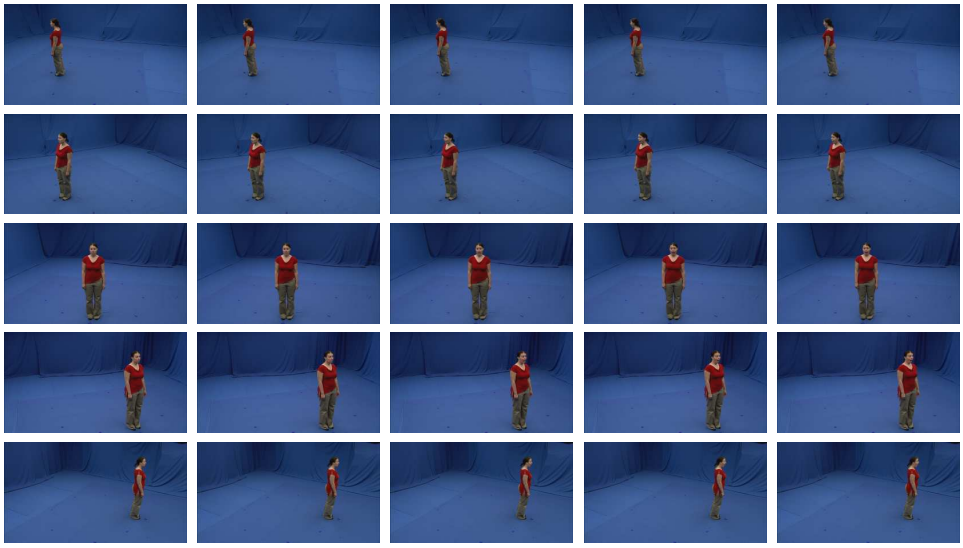


Figure 7: *At the five top rows, five frames depicting the same person executing the facial expression of surprise taken from five different cameras is depicted. The cropped facial region is depicted for the frames produced by the camera that captures the frontal facial view.*

Figure 8: *Cropped facial regions corresponding to the frames produced by the camera that captures the frontal facial view (middle row in Figure 7).*

case of a studio setting where the foreground can be reliably and consistently extracted from a known fixed background. However, SFS only provides an upper bound on the volume of the scene. Concavities that are occluded in silhouettes are not reconstructed, appearance is not matched across images and phantom false-positive volumes can occur that are consistent with the image silhouettes. Figure 10 shows the reconstructed visual-hull from multiple silhouette images in which surface concavities are not represented and phantom volumes are incorporated into the recovered surface.

In our case, the scene is finally reconstructed as the surface within the visual-hull that passes through the surface features while maximizing the consistency in appearance between views [21, 22]. Once the extent of the scene is defined by reconstructing the visual-hull, surface features are matched

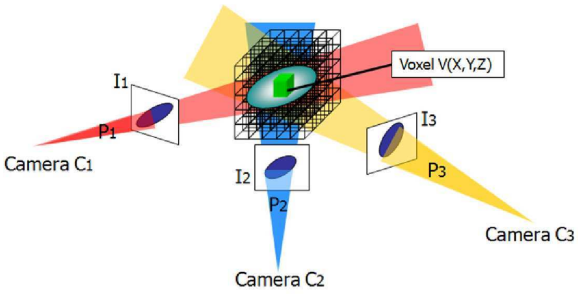


Figure 9: Shape-from-Silhouette.

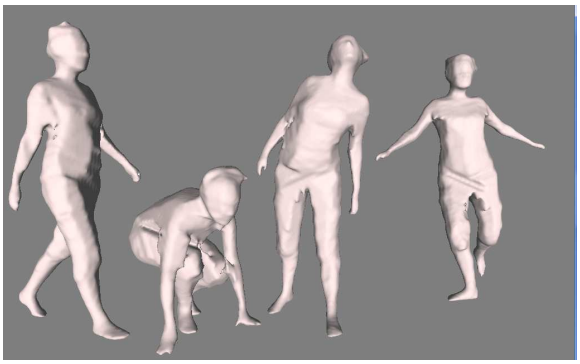
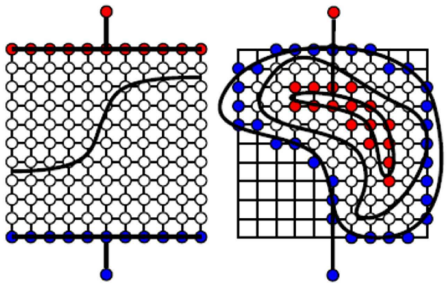


Figure 10: Results by SFS.

between views to derive constraints on the location of the scene surface. Surface features correspond to local discontinuities in the surface appearance and candidate features are extracted in the camera images using a Canny-Deriche edge detector. Each feature contour in an image is then matched with the appearance in an adjacent camera view. The connected set of pixel correspondences are then derived in the adjacent view that maximizes the image correlation for the feature contour. Correspondence is verified by enforcing left-right consistency between views such that a feature pixel in one camera is required to match a feature pixel in an adjacent camera with a reciprocal correspondence.

However, the feature reconstruction provides only a sparse set of 3D line segments that potentially lie on the scene surface. Dense surface reconstruction is then performed inside the volume defined by the visual-hull. We use a global optimization approach by discretizing the volume and treating reconstruction as a maximum-flow/minimum-cut problem on a graph defined in the volume. Surface reconstruction as a network flow problem on a graph is illustrated in Figure 11. Each voxel forms a node in the graph with adjacent voxels connected by graph edges between a source (blue) and sink (red). Edges are weighted by a cost defined by the consistency in appearance between camera images. The maximum flow on the graph saturates the set of edges where the cost is minimized and the consistency is maximized. The final surface can then be extracted as the set of saturated edges cutting the graph. Efficient optimization methods exist using graph-cuts,



(a) Graph for a 2D image

(b) Graph for a 3D structure

Figure 11: Graph-Cut.

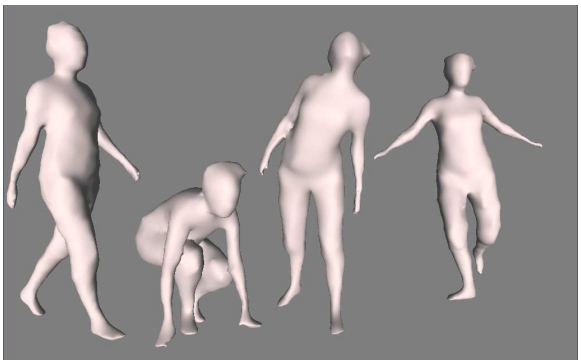


Figure 12: Results of surface refinement.

providing the global optimum that maximizes the correlation between views on the final surface [5].

Finally, the surface for the scene is extracted from the volume reconstruction as a triangulated mesh. Mesh vertices are derived to sub-voxel accuracy using a local search to maximize image consistency across all visible cameras. Figure 12 shows the result of surface reconstruction.

3 Database application example: View-invariant human activity recognition from multi-view video

In this section we use a part of the database to evaluate the performance of a view-independent activity recognition algorithm that resulted by extending a single-view method published in [9], thus, providing an example application of the database.

3.1 View-invariant human movement representation

We assume that a convergent camera setup of Q cameras, similar to the one described in section 2.1, is used to create human movement videos and that the procedure described in section 2.3 is applied to create the respective body mask ROI sequences.

The mask sequences that depict a person executing the same

movement more than once are manually segmented in the temporal dimension to their constituting single movement videos. The centroid of the body posture at each image is computed and each movement video is aligned with respect to the body centroid and rescaled to the same size of $W \times H$ pixels, where W is the width and H is the height of the image. Let us denote the aligned and rescaled multi-view movement video as $\{\mathbf{x}_{i,q}^d\}_{i=1,\ldots,L,q=1,\ldots,Q}$ where $\mathbf{x}_{i,q}^d \in \Re^F$ is a vector of length f ($F = WH$) created by scanning column-wise the i-th image of the q-th video stream, L is the length of the movement video in frames, and the superscript d denotes the 3D posture captured from the d-th viewing angle, and should not be confused with the camera index q. All the single-view posture vectors referring to the same frame instance are concatenated to produce the so called multi-view posture vector $\mathbf{x}_i^d \in \Re^{FQ}$:

$$\mathbf{x}_i^d = \left[\mathbf{x}_{i,1}^{d\ T}, \ldots, \mathbf{x}_{i,Q}^{(d+Q)_Q\ T}\right]^T , \quad (1)$$

where $()_Q$ denotes the modulo Q operator and d is the (unknown) view index for the first camera. Obviously, since the subject may move freely within the view volume, the problem of view correspondence, i.e., which view of the 3D human body, denoted by the index d, is captured by the q-th camera, does not allow us to train a movement classifier using directly the multi-view posture vectors. To solve this problem, we observe that all Q possible configurations, $\mathbf{x}_i^d, d = 1, \ldots, Q$, of a multi-view posture vector can be obtained by block circularly shifting its elements, with a block corresponding to a single-view posture vector. By denoting the elements of a multi-view posture vector as $x_i^l(n), n = 1, \ldots, N$, ($N = QF$), this circular shifting can be formally written as $x_i^k(n) = x_i^l((n - |k-l|F)_N)$, where $|\cdot|$ denotes absolute value. Based on this observation, the view correspondence problem can be implicitly solved using the magnitude of the discrete Fourier transform (DFT) coefficients of the multi-view posture vector:

$$\tilde{x}_i(k) = |\sum_{n=0}^{N-1} x_i^d(n) \exp^{-j2\pi kn}|,\ k = 1, \ldots, N , \quad (2)$$

Concatenating the above coefficients to form the vector $\tilde{\mathbf{x}}_i$, a view-invariant representation of the posture is constructed. Thus, a movement is described with the respective sequence $\{\tilde{\mathbf{x}}_i\}_{i=1,\ldots,L_i}$, where the superscript d has been dropped as this representation is view-invariant.

3.2 Human movement recognition

Let $\mathcal{U}$ be an annotated database of multi-view movement sequences $\{\tilde{\mathbf{x}}_{i,j}, y_j\}$ belonging to one of R classes, where the vector $\tilde{\mathbf{x}}_{i,j}$, described in the previous paragraph, represents the i-th frame of the j-th movement sequence and $y_j \in [1, \ldots, R]$ is its label.

Without considering the labelling information, the fuzzy c-means (FCM) algorithm is used to compute C centroids, $\mathbf{v}_c, c = 1, \ldots, C$ and partition the data to C classes, where the number of centroids C and the fuzzification parameter m are assumed known. Using the computed centroids, fuzzy vector quantization (FVQ) is applied to quantize the posture vectors $\boldsymbol{\phi}_{i,j} \in \Re^C$, $\boldsymbol{\phi}_{i,j} = [\phi_{c,i,j}]$, where,

$$\phi_{c,i,j} = \frac{(\parallel \tilde{\mathbf{x}}_{i,j} - \mathbf{v}_c \parallel_2)^{\frac{2}{1-m}}}{\sum_{j=1}^{C}(\parallel \tilde{\mathbf{x}}_{i,j} - \mathbf{v}_j \parallel_2)^{\frac{2}{1-m}}} . \quad (3)$$

The j-th sequence is then represented by the arithmetic mean of its quantized postures,

$$\mathbf{s}_j = \frac{1}{L_j}\sum_{i=1}^{L_j} \boldsymbol{\phi}_{i,j} . \quad (4)$$

The labelling information can be further exploited to reduce the dimensionality of the feature vectors using linear discriminant analysis (LDA. Assuming that $\boldsymbol{\Psi} \in \Re^{C \times R-1}$ is the projection matrix computed using LDA the final representation of the video is $\mathbf{z}_j = \boldsymbol{\Psi}^T \mathbf{s}_j$. The rth movement type can then be represented by the mean of all feature vectors belonging to this movement type, i.e.,

$$\boldsymbol{\zeta}^{(r)} = \frac{1}{O_r}\sum_{\mathbf{z}_j \in \mathcal{U}_r} \mathbf{z}_j . \quad (5)$$

where $\mathcal{U}_r$ denotes the set of the videos belonging to the r-th class and O_r is the cardinality of the r-th class. During recognition, the feature vector of the test video is retrieved, and a cosine similarity value between the test feature vector and each movement prototype vector is computed. Subsequently, the test video is classified to the class represented by the prototype that produced the maximum cosine value.

3.3 Experimental evaluation of the method on the i3DPost database

To evaluate the performance of the proposed methods 40 multi-view videos were selected from the i3DPost database. More specifically, the videos of all eight persons for walk (wk), run (rn), jump in place (jp or jump1), jump forward (jf or jump2) and bend (bd) were utilized. The above movements are described in detail in section 2.2.1.

For the evaluation of the algorithm the leave-one-out-cross-validation (LOOCV) procedure is used. At each LOOCV cycle, the multi-view video of a specific person executing a specific movement is used as a test video, and the rest of the videos in the database form the training set. The training videos are preprocessed as described in section 3.1 to produce a set of movement sequences $\{\tilde{\mathbf{x}}_{i,j}, y_j\}$, and subsequently to learn the movement prototypes $\boldsymbol{\zeta}^{(r)}$ as described in section 3.2. During testing, if necessary, the test video is manually segmented to its constituting single period movement videos and each of them is classified. The final decision for the entire test video is taken using majority voting.

Similar to [9], the LOOCV procedure was combined with the global-to-local search strategy to identify the optimal parameters, which in our case are $C = 20$ and $m = 1.1$.

For these values, as shown in Table 2, only two videos were misclassified, while other two remained unclassified, giving a 90% correct recognition rate.

	wk	rn	jf	jp	bd	–
wk	8					
rn		7				1
jf			7			1
jp			1	6	1	
bd					8	

Table 2: *Confusion matrix for the five movements from the i3DPost database. The last column corresponds to videos that remained unclassified.*

4 Conclusions

It is expected that human action recognition algorithms utilizing multi-view/3D videos will offer better recognition rates than methods that use single-view videos. To facilitate the development of multi-view algorithms we have created a publicly available multi-view and 3D human action/interaction database. This database contains videos of eight persons and twelve actions captured from eight high resolution cameras. In addition, a sequence that contains the basic facial expressions is recorded for each person, providing a total of 104 multi-view videos. To increase the applicability of the database, each multi-view video has been preprocessed to provide the respective binary mask sequence, posture ROI sequences as well as a 3D body mesh per frame instance. Our hope is that the database will serve as a testbed for the development, evaluation and/or comparison of human action recognition algorithms based on multi-view/3D videos.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 211471 (i3DPost).

References

[1] J. K. Aggarwal and S. Park. Human motion: modeling and recognition of actions and interactions. In *2nd International Symposium on 3D Data Processing, Visualization and Transmission*, pages 640–647, Thessaloniki, Greece, September 2004.

[2] M. Ahmad and S. W. Lee. HMM-based human action recognition using multiview image sequences. In *18th Int, Conf. Pattern Recognition*, volume 1, pages 263–266, Hong Kong, China, August 2006.

[3] A. F. Bobick. Movement, activity and action: The role of knowledge in the perception of motion. *Philosophical Transactions: Biological Sciences*, 352(1358):1257–1265, 1997.

[4] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Machine Intell.*, 23(3):257–267, March 2001.

[5] Y. Boykov and V. Kolmogorov. Computing geodesics and minimal surfaces via graph cuts. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, Washington, DC, USA, 2003. IEEE Computer Society.

[6] `http://mocap.cs.cmu.edu`.

[7] `http://gesturedb.korea.ac.kr`.

[8] M. Flierl and B. Girod. Multiview video compression. *IEEE Signal Processing Mag.*, 24(6):66–76, November 2007.

[9] N. Gkalelis, A. Tefas, and I. Pitas. Combining fuzzy vector quantization with linear discriminant analysis for continuous human movement recognition. *IEEE Trans. Circuits Syst. Video Technol.*, 18(11):1511–1521, November 2008.

[10] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Trans. Pattern Anal. Machine Intell.*, 29(12):2247–2253, December 2007.

[11] R.D. Green and L. Guan. Quantifying and recognizing human movement patterns from monocular video images-part I: A new framework for modeling human motion. *IEEE Trans. Circuits Syst. Video Technol.*, 14(2):179–190, February 2004.

[12] F. Huang and G. Xu. Action recognition unrestricted by location and viewpoint variation. In *Proc. IEEE 8th Int. Conf. on Computer and Information Technology Workshops*, volume 00, pages 433–438, Sydney, Australia, July 2008.

[13] F.Y. Huang and G.Y. Xu. Viewpoint insensitive action recognition using envelop shape. In *8th Asian Conf. on Computer Vision*, pages II: 477–486, Tokyo, Japan, November 2007.

[14] B.-W. Hwang, S. Kim, and S.-W. Lee. A full-body gesture database for human gesture analysis. *International Journal of Pattern Recognition and Artificial Intelligence*, 21(6):1069–1084.

[15] I. N. Junejo, E. Dexter, I. Laptev, and P. Pérez. Cross-view action recognition from temporal self-similarities. *Computer Vision ECCV 2008*, pages 293–306, 2008.

[16] A.S. Ogale, A. Karapurkar, and Y. Aloimonos. View-invariant modeling and recognition of human actions using grammars. In *Workshop on Dynamical Vision at 10th IEEE Int. Conf. Computer Vision*, Beijing, China, October 2005.

[17] V. Parameswaran and R. Chellappa. Human action-recognition using mutual invariants. *Comput. Vis. Image Underst.*, 98(2):295–325, 2005.

[18] V. Parameswaran and R. Chellappa. View invariance for human action recognition. *Int. J. Comput. Vision*, 66(1):83–101, 2006.

[19] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *17th International Conference on Pattern Recognition*, volume 3, pages 32–36, Cambridge, UK, August 2004.

[20] C. Shen, C. Zhang, and S. Fels. A multi-camera surveillance system that estimates quality-of-view measurement. In *Proc. IEEE Int. Conf. Image Processing*, volume 3, pages III–193 – III–196, San Antonio, Texas, USA, September 2007.

[21] J. Starck and A. Hilton. Surface capture for performance-based animation. *IEEE Comput. Graph. Appl.*, 27(3):21–31, May 2007.

[22] J. Starck, A. Hilton, and G. Miller. Volumetric stereo with silhouette and feature constraints. In *British Machine Vision Conference*, September 2006.

[23] P. Turaga, R. Chellappa, V.S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Trans. Circuits Syst. Video Technol.*, 18(11):1473–1488, November 2008.

[24] Y. Wang, K. Huang, and T. Tan. Human activity recognition based on $\mathcal{R}$ transform. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8, June 2007.

[25] Y. Wang, K. Huang, and T. Tan. Multi-view gymnastic activity recognition with fused hmm. In *8th Asian Conference on Computer Vision*, volume Part I, pages 667–677, Tokyo, Japan, November 2007.

[26] D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3d exemplars. In *IEEE 11th International Conference on Computer Vision*, pages 1–7, Rio de Janeiro, Brazil, October 2007.

[27] D. Weinland, R. Ronfard, and E. Boyer. Motion history volumes for free viewpoint action recognition. In *EEE International Workshop on modeling People and Human Interaction*, October 2005.

[28] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Comput. Vis. Image Underst.*, 104(2-3):249–257, Nov-Dec 2006.

[29] Y. Yang, A. Hao, and Q. Zhao. View-invariant action recognition using interest points. In *1st ACM international conference on Multimedia information retrieval*, pages 305–312, 2008.