The Nucleus ***55***, *No. 3 (2018) 138-148*

www.thenucleuspak.org.pk

The Nucleus

ISSN 0029-5698 (Print)

ISSN 2306-6539 (Online)

Spatio - Temporal RGBD Cuboids Feature for Human Activity Recognition

H.A. Sial[1], M.H. Yousaf[2] and F. Hussain[2*]

[1]*Department of Computer Science & Engineering, University of Engineering & Technology Lahore, Narowal Campus, Pakistan*

[2]*Department of Computer Engineering, University of Engineering & Technology Taxila, Pakistan*

ARTICLE INFO

Article history :

Received : 09 March, 2018

Accepted : 12 November 2018

Published : 13 November, 2018

Keywords:

Human activity recognition

Depth sensor,

Bag of words

Interest points.

ABSTRACT

Human activity recognition is one of the promising research areas in the domain of computer vision. Color sensor cameras are frequently used in the literature for human activity recognition systems. These cameras map 4D real-world activities to 3D digital space by discarding important depth information. Due to the elimination of depth information, the achieved results exhibit degraded performance. Therefore, this research work presents a robust approach to recognize a human activity by using both the aligned RGB and the depth channels to form a combined RGBD. Furthermore, in order to handle the occlusion and background challenges in the RGB domain, Spatial-Temporal Interest Point (STIP) based scheme is employed to deal with both RGB and depth channels. Moreover, the proposed scheme only extracts the interest points from depth video (D-STIP) such that the identical interest points are used to extract the cuboid descriptors from RGB (RGB-DESC) and depth (D-DESC) channels. Finally, a concatenated feature vector, comprising features from both channels is passed to exploit a bag of visual words scheme for human activity recognition. The proposed combined RGBD features based approach has been tested on the challenging MSR activity dataset to show the improved capability of combined approach over a single channel approach.

1. Introduction

Human Activity Recognition has become an actively researched area for a wide range of real-world applications. The objective of activity recognition is to detect and identify the activity in real-world problems or in the video dataset. Activities can be comprised of day-life activities, e.g. talking on the phone, eating, drinking, playing guitar, using a laptop and walking, etc. One may find its diversified applications in different fields of life, e.g., surveillance system, health and medical fields, sports, human-computer interaction, intelligent home and office environments, etc. [1]. It is evident from previous research that activity recognition is mostly done using color sensor cameras only [2]. However, color sensor cameras only consider *x-y-t* dimensional volume. On the other hand, real-world activities are four-dimensional *x- y- z- t*. Therefore, color camera maps 4D real-world activities to 3D digital space by discarding vital depth information. The non-existence of depth information often leads to misclassification and misinterpretation of visual scenes.

The recent development of depth sensors-based cameras like Microsoft Kinect empowered researchers to solve the deficiency of depth information. Despite the above-mentioned issues, color sensor cameras are more sensitive to illumination, color and occlusion problems, making recognition task more problematic [2]. The depth sensor offers many advantages over the color sensor, such as it can work in poor lighting conditions and provides 3D structural information. These advantages can enable the recognition system to recognize patient or elderly people's activities continuously round the clock. Microsoft Kinect also offers the 3D joint position and orientation information in human skeletons which makes recognition even simpler, but this powerful tool has certain limitations, since skeletons are only accurately modeled when an object is facing the camera in an upright position. Skeleton estimation can hardly work when the body is occluded or not facing towards the camera, e.g., person lying on the sofa, monitoring human activities in surveillance [2].

In the RGB domain, spatial-temporal interest points (STIP) can provide one of the robust ways to handle occultation, illumination and cluttered environment challenges [1]. STIP represents an abrupt change in video and this is a most informative point to model distinctive activity. Considering the availability of both RGB and depth channel information, the proposed research work is focused on handlingabove mentioned discrepancies in an efficient way.

Three principle questions are answered in this research:

1. Whether the algorithms should be designed to perform a recognition task, using a single channel or multiple channels for improved results?
2. Whether the interest points should be detected independently from each channel or the depth channel interest points can be merged into RGB data?
3. What is the experimental cost if we use multiple channels instead of a single channel?

Consequently, both channels (RGB as well as depth) features are combined and a bag of visual words are used to recognize a human activity. RGBD bag of words can recognize human activities without any dependence on skeleton tracking or preprocessing like human posture

*Corresponding author : fawad.hussain@uettaxila.edu.pk

segmentation or motion tracking. To show the viability of our approach, it has been tested on the challenging MSR activity dataset [3]. The experimentation provides promising results in comparison with any state-of-the-art algorithm, based on a low level or high-level features. Furthermore, the proposed framework is more widely applicable on RGB-Depth senor based human activity recognition, e.g., in surveillance system where the camera is mounted higher than the human body and both color and depth channel information are available and aligned.

The rest of this paper is organized as follows: Section 2 provides a critical review of the related work in the domain of human activity recognition. Section 3 presents a detailed description of the feature extraction and the description process from RGB and depth channels to recognize human activities. Section 4describes the results and analysis of the proposed system. Finally, Section 5 concludes the article with some probable future direction.

2. Related Work

The objective of Human Activity Recognition is to recognize and classify real-world activities in videos automatically. Vision-based recognition system can be categorized into four levels, i.e., gesture, interaction, action, and group activities [4]. Despite extensive research in this field, it is still challenging due to environmental constraints such as moving background, occlusion and viewpoint orientations. It also has challenges to resolve illumination problems, intraclass variations and due to the availability of sufficient training data [4]. An efficient and robust recognition system should be able to handle these variations. Researchers have proposed many state-of-the-art algorithms to recognize human activities in RGB, depth and combined RGBD channels. Both RGB and depth sensor-based approaches can be categorized in STIP local and global representation approaches.

2.2 Color Sensor-Based Approaches

Various human activity recognition approaches have been proposed in the literature using color sensor camera due to its low cost and availability. Space-time interest points (STIP's) represent abrupt changes in the video such that the interest points are distinct for each activity class and are most informative. Many 2D STIP based approaches [5-8] have been extended in 3D to recognize a human activity. Most widely used STIP detectors in RGB channel are Harris 3D [9], Dollar cuboid detector [10], Hessian detector [11] and dense sampling [12]. After detection of spatio-temporal interest points in the video, these points are needed to be represented by distinctive descriptors. The descriptors should be invariant to spatial-temporal scales, environmental constraints, and image rotation. Prominent STIP descriptors are Dollar cuboids [10], HOG/HOF [13], HOG 3D [14] SURF 3D [11, 15] and SIFT 3D [16]. Wang et al. [17] introduced motion-lets based on a histogram of energy and histogram of gradients. Motion-lets represent both appearance and motion information.

Researchers have also proposed some global representation approaches that can be designed to work on a specific recognition application. Bobick et al. [18] were first to use this approach. They characterized an activity with two image pairs, called Motion history image (MHI) and motion energy image (MEI). Ke et al. [19] used an RGB segmentation approach to model ROI such that the similar color pixels are clustered together to find an action model shape. Sheikh et al. [20] represented the human body by its joints positions and used joint trajectory to symbolize an activity. Veeraraghavan et al. [21] used dynamic time wrapping (DTW) in an activity recognition task by modeling speed variations of all the activities. Hidden Markov Model (HMM) techniques are also widely used in the literature [22]. In HMM an activity is represented by unseen states. Therefore, the system assumes human in one state and considers the probability of transition between states.

Color sensor-based approaches are sensitive to illumination changes, color and background motion in activity recognition task, they discard vital depth cue in activity recognition task. In-depthchannel, one can better control background motion and illumination problems. Proposed system utilizes both channels, which is advantageous and outperforms existing individual channel-based methods.

2.3 Depth Sensor Based Approaches

Depth human activity recognition is a comparatively new and exciting field in computer vision. It has gained researchers attraction due to new development of depth sensor camera like Microsoft Kinect. Recently, some depth STIP detectors and descriptors have been implemented in this exciting research field. Cheng et al. [23] introduced a new descriptor, named as comparative coding descriptor (CCD), which is an extension of local binary partition (LBP) [24]. In a 3D domain, the $3 \times 3 \times 3$ region is considered around an interest point, and difference of depth value between the central point and 26 neighboring points is coded. Yang et al. [25] used local depth pattern (LDP) to represent the depth data, which is a 2D descriptor, where the local region around interest point is divided into $nx \times ny$ sub regions, such that in each sub cell the average depth value is computed. Wang et al. [26] introduced Local occupancy patterns (LOP), based on 3D cuboids around joint points in a skeleton. Wang et al. [27] introduced semi local feature called Random Occupancy Pattern (ROP), which is based on the integral image and sparse coding scheme and it can handle occlusion challenges. Xia and Aggarwal [2] introduced Depth spatial temporal interest point and descriptor to recognize an activity in depth data to handle the environmental challenges.

Global representations are more efficient in-depth data, but they have limited applications, since skeleton information is not always accessible. Wanqing et al. [3] used a concept of a bag of 3D points to recognize human activity in depth map, in which 3D silhouette of human is represented by few sample points. Space-time occupancy patents (STOP) was introduced by Vieira et al. [28], in STOP each depth map sequence is represented by the4D grid to preserve both spatial and temporal information. Yang et al. [29] introduced depth motion map (DMM), in this algorithm each depth frame is projected on 3 Cartesian planes to get projected depth map. Histogram of gradient is applied on these binary maps to represent action by features [30]. Recently Histogram of oriented 4D Normal's (HOND 4D) was presented [31], HOND 4D represent shape and motion information concurrently by making a histogram that represents surface normal distribution in depth and spatial coordinates. Seidenari et al. [32] introduced a new recognition system based on skeleton joints positions; in this task, they represented joints on anew coordinate system which makes joints invariant to rotation and transition. Zanfir et al. [33] introduced kinematic descriptor for activity recognition system, the kinematic system captures both joints position information and their velocity and acceleration parameters.

Depth data has certain advantages in handling viewpoint invariance, scale invariance and lighting condition; it can work in poor lighting condition. Most of depth channel activity recognition research has been done using global skeleton-based methods with the certain assumption regarding camera position which makes these algorithms to work in a limited range of applications. In surveillance systems, the camera is always mounted above the human body, and in that scenario, it's hard to get perfect skeletons. Our approach is not dependent on skeleton extraction and can work flawlessly when both channels information are readily obtainable.

2.3 RGBD Approaches

In this kind of approaches, both RGB and depth channels are considered for feature extraction. This is a relatively new area of research in computer vision and provides better results as compared to single channel approaches. Sung et al. [34] introduced hierarchical Maximum Entropy Markov Model (MEMM). In MEMM, an activity is subdivided into smaller activities and considers two layers HMM approach on these activities for classification purpose. RGB depth sensor camera, Kinect is used in this approach to get the skeleton information. Body pose features, hand position and motion information are used to represent an activity feature vector. Ni et al. [35] extended the idea of motion history images (MHI) for RGB depth videos [18]. They introduced two MHI in-depth data: forward depth motion history image (FDMHI) encodes the forward motion (an increase of depth) and backward-DMHI (BDMHI) encodes the backward motion (decrease of depth). Koppula et al. [36] modeled human interaction activities with objects using object nodes and human skeleton sub activity information. Object node represents object position and its temporal movement using SIFT [7] tracker. On the other hand, the sub activity node represents the human activity information using the skeleton, obtained in the RGB depth video.

All existing RGBD approaches work on global information. Furthermore, most of these algorithms/ approaches depend on the skeleton information. However, the skeleton information is not always accessible. To the best of our knowledge, the proposed approach is the first RGBD spatial-temporal interest point based method that exploits the advantages of both channels and out performs individual channel-based approaches on a challenging MSR activity dataset. STIP based methods have shown their significance to handle the occlusion and background clutter problems in the RGB domain. Our method detects these interest points from the depth channel. The proposed approach efficiently handles the background motion in the activity recognition and enables us to remove many false interest points. Furthermore, the proposed method is not dependent on the skeleton information. Consequently, it can be applied in real-worlds scenarios when both color and depth channel information are available and synchronized.

3. Proposed RGBD Approach for Activity Recognition

Proposed human activity recognition algorithm is designed to work when both RGB and depth channels are available and aligned. Fig. 1 shows the complete framework in which the depth spatial-temporal interest points (D-STIP) from are first extracted from the depth data and then mapped to the RGB channel. Depth descriptors (D-DESC) and RGB descriptors (RGB-DESC) are computed using the depth interest points only. Each descriptor's length is reduced using principal component analysis. Reduced descriptors are concatenated, and the bag of visual word approach is used to make a feature vector. Bag of visual word approach fastens recognition task by converting an activity Np Descriptors to 1D feature vector. Support Vector Machine (SVM) is used to classify an activity. Each of the above-mentioned steps is explained in the subsequent sections.

3.1 Extraction of Depth Features

The proposed system for depth feature extraction is based on Xia and Aggarwal [2] research work. Proposed work represents each human activity with local cuboids features, based on the bag of visual words. It can also work in the scenario, where no human skeletons are available. Skelton extraction is unlikely to obtain when the human body is occluded or touches the background.

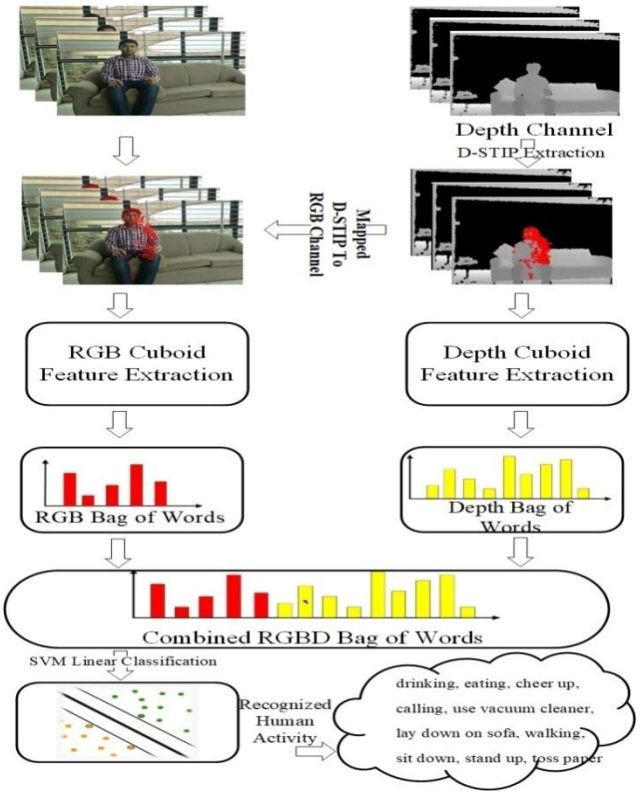

Fig. 1: The framework of the proposed approach.

Main steps of Depth Feature Extraction are as follow:

- Extraction of depth STIP (D-STIP) from depth video dataset using Gabor filter.
- Extraction of Depth Descriptor (D-DESC) centered at D-STIP, having cuboids of variable size based on the depth value.
- Represent multiple local D-DESC as a feature vector using the bag of visual words approach (BOVW).

3.1.1 D-STIP extraction

In most of the earlier research, regarding the interest point extraction using the RGB channel only, RGB feature extraction algorithm cannot provide satisfactory results on the depth data. The primary reason is that the depth data has a different texture and various nature of noise. In any feature extraction algorithm, the main task is to compute the repose function of filters at each pixel location. D-STIP extraction framework is based on Gaussian filter along the spatial dimension (x, y) and 1D complex separable Gabor filter along the temporal (t) direction.

First, each frame in the depth video is convolved with Gaussian smoothing filter independently, as evident from equation (1). This is a preprocessing step to remove sharp changes.

$$D_s(x, y, t) = D(x, y, t) * G(x, y) \quad (1)$$

$D(x, y, t)$ is the input video volume such that the each video volume frame is convolved (*) with a 2D Gaussian smoothing kernel $G(x, y)$, where $G(x, y)$ is given in equation (2). D_s is the smoothed output video volume.

$$G(x, y) = \frac{1}{2\pi {\sigma_d}^2} e^{-\frac{x^2+y^2}{2{\sigma_d}^2}} \quad (2)$$

σ_d is standard deviation of the filter that controls the spatial scale along x and y. After applying the spatial Gaussian filtering, D_s is convolved with 1D complex Gabor filter,0 as described in equation (3).

$$Dst(x, y, t) = Ds(x, y, t) * h(t|\, \tau_d) \quad (3)$$

$Dst(x, y, t)$ represent the output video volume after 1D convolution in temporal domain with $h(t|\, \tau_d)$ see equation (4).

$$h(t|\, \tau_d) = e^{-t^2/\tau^2} . e^{2\pi\omega\tau_d} \quad (4)$$

τ_d represent the temporal scale of the detector at which the depth interest points are detected. Our response function is computed using $\omega_d = 0.6/\tau_d$. $h(t|\, \tau_d)$ can be represented in the form of even and odd separate filter hev and $hod$and can be applied separately and combined later to reduce the computational complexity. Values of hev and hod are mentioned below in equation (5) and (6), respectively.

$$hev = -cos(2\pi t\omega_d) e^{-t^2/{\tau_d}^2} \quad (5)$$

$$hod = -sin(2\pi t\omega_d) e^{-t^2/{\tau_d}^2} \quad (6)$$

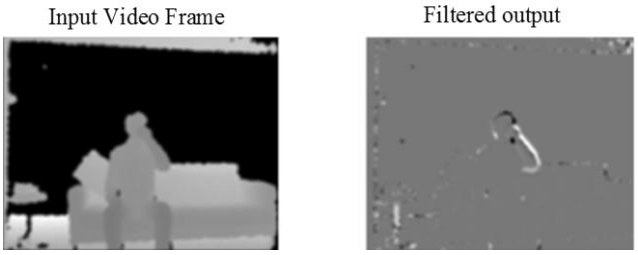

Fig. 2: Filter response of Gaussian and Gabor filters.

Fig. 2 shows the filter response of Gaussian and Gabor Filters. In the Right side of the image, the brighter white portion shows the change in the video. Gabor filter provides motion pixels in the temporal domain. In the RGB domain, the noise is generally removed by using smoothing filters, but it is not possible in the depth domain due to the different nature of noise. Consequently, to remove depth channel noise, we have to make a correction function instead of any filter response. Depth noise can be categorized into three categories:

- The first category is similar to the color channel noise, and the major cause for this type of noise is a sensor fault. This type of noise is distributed throughout the video frames and is relatively small in magnitude.

- The second category of noise occurs in the depth video around the boundary of objects; this noise moves back and forth from the foreground depth value to the background depth value. The magnitude of this kind of noise is relatively large.
- The thirdcategory of noise is "hole," caused due to the fast movement or incorrectly assigned depth pixel value by any reflecting material or any other random reason. This noise can usually be seen in the middle of static objects.

Fig. 3 shows the temporal values of true motion pixels and boundary pixels. The first type of noise can be removed using a Gaussian smoothing filter, while second and third categories cannot be removed using any filter response. Fig. 3 shows noise around the boundary pixels, fluctuating around the mean value.

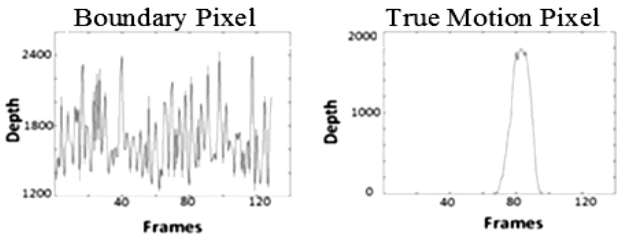

Fig. 3: Temporal values of boundary and true motion pixels.

The magnitude of this type of noise is comparatively large then real movements, so any filter response on this type of noise will affect the real movement. Flips in the noise are much faster than areal movement. To remove such noise, the average duration of flips is calculated and used as a correction function, as mentioned in equation (7).

$$Cr(x,y) = \frac{\sum_{i=1}^{n_{fp}} \delta t_i(x,y)}{n_{fp}(x,y)} \quad (7)$$

Here $n_{fp}(x,y)$ is the total no of fluctuations in the whole video at location (x,y)while$\delta t_i(x,y)$ calculates the time for ith flip. Similarly, the numbers of flips are defined as a signal crossing around the mean value $(d(t)max + d(t)min)/2$. Correction function is a 2D image, where each pixel represents signal-noise ratio at (x,y) during the whole video and it will have larger values at real movement pixels as compared to noisy pixels. Therefore, we define a threshold to distinguish between real movement and noise. Fig. 4 shows the correction function for drinking activity. This correction function is multiplied with each video frame to suppress noise values.

Overall response function can be written, as given in equation (8):

$$R_d = (D * G * hev\ o\ Cr)^2 + (DG *)^2 \quad (8)$$

In which o represents the pixel multiplication of correction function with filtered video volume. D-STIP's are selected by taking local maxima of a pixel in spatial and temporal domains.

Fig. 4: Correction function response.

Top Np points can be selected from the final response function. Fig. 5 shows the interest points for corrected and non-corrected video volume. Interest points are shown only on one frame of drinking activity for visual purpose only.

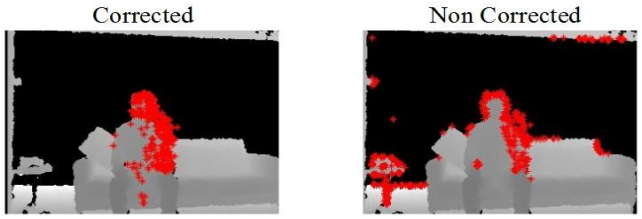

Fig. 5: Detected D-STIP for corrected and non-corrected video.

3.1.2 Depth descriptor (D-DESC) extraction

3D spatio-temporal cuboids are extracted around the interest points based on the depth value, the spatial scale σ_d and the temporal scaleτ_d. As any object in the image appears smaller at a farther distance, the descriptor should be designed considering its depth value. In our implementation, we defined a cuboid size that is inversely proportional to the interest point depth according to equation (9) and is given as:

$$\Delta x^{(i)} = \Delta y^{(i)} = \sigma_d \frac{L}{d^i} \quad (9)$$

σ_d is the spatial scale of Gaussian filter used in feature extraction, L is the entity scalar called support region and d^i is the minimum non-zero depth value in $2\tau_d$ window around the interest point. Reason for choosing $2\tau_d$ window is that sometimes the interest point is detected at the edge or boundary pixel. In order to match with the real-world size object, we assign a smaller cuboid size to the farther distant object and a larger size to the closer objects. Consequently, the temporal direction of cuboid is defined according to the equation (10) and is given below:

$$\Delta t_i = 2\tau_d \quad (10)$$

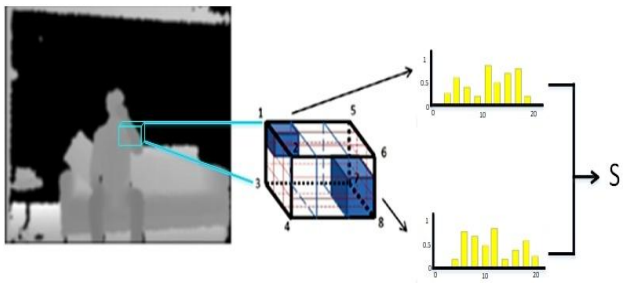

Fig. 6: Depth cuboid extraction process.

To encode shape information, the D-DESC implementation is based on the self-similarity in the cuboid region. Each cuboid of size $(px, py, pz)$is divided in to $n_{xy} \times n_{xy} \times n_t$ sub blocks. At the same time, the border pixels are dropped to obtain integer no of pixel values. Each local region is defined from $1 \times 1 \times 1$ to $n_{xy} \times n_{xy}$ xn_t number of sub blocks. In each local region, depth values histogram is computed and normalized. Histogram is computed by quantizing depth values in M bins. Range of quantizer varies from minimum d_{min} to maximum d_{max} depth value with equal step size $\Delta d = d_{max} - d_{min}$ / M.Histogram $h(n)$is defined using equation (11):

$$h(n)\begin{cases} 1, \ if \ (n-1)\Delta d + 1 \le D(x,y,t) \le n\Delta d \\ 0, \ else \end{cases} \quad (11)$$

Two local region histogramshp and hqsimilarity score is computed using Bhattacharyya distance, as mentioned in equation (12) below:

$$S(p,q) = \sum_{n=1}^{M} \sqrt{hp \ hq} \quad (12)$$

Similarly, all histogram similarity scores are concatenated to form a feature vector. Fig. 6 explains the feature extraction process. Length of the feature vector is not depended on cuboid size; it only depends on n_{xy} andn_t. Total number of local regions N_b that are generated by varying block from $1 \times 1 \times 1$ to $n_{xy} \times n_{xy} \times n_t$ is shown below in equation (13):

$$N_b = \frac{n_{xy}(n_{xy}+1)(2n_{xy}+1)}{6} \times \frac{n_t(nt+1)}{2} \quad (13)$$

Total length l_d of D-DESC is defined as $l_d = C_{N_b}^2$. Depth extraction process will generate $Np \times l_d$ matrix for Np Depth interest point (D-STIP). Local Depth feature extracts motion and shape information that is invariant to occlusion, clutter environment, scale and spatial temporal shift in video.

3.2 Extraction of RGB Features

RGB Feature extraction is based on the research work of Dollar et al. [10]. RGB data has better texture as compared to depth map and simple gradient based descriptors. It can provide promising results in an activity recognition system. Furthermore, RGB data also has a different nature of noise which can easily be removed with smoothing filter. Moreover, extraction of RGB feature is also based on the local low-level features, and its methodology is quite similar to the depth feature extraction. RGB-STIP's are only required when we consider the RGB channel independently for activity recognition. The RGB-STIP extraction process is also based on the separable Gaussian smoothing filter and 1D Gabor temporal Quadrature filter pair hev and hod [37]. Overall response function for color feature extraction can be described using equation (14):

$$R_c = (C * G * hev)^2 + (C * G \ hod)^2 \quad (14)$$

C represents the color video frames; G is the 2D Gaussian smoothing filter convolved (*) with the RGB channel data. The cuboid detector has a strong response against the periodic motion. The response function gives a maximum response to the activities like walking. Regional maxima of 26 connected neighbors in $3 \times 3 \times 3$is found to get desired interest points.Fig.7 shows the detected RGB-STIP on drinking activity by projecting all interest point on (x, y) plane. These points are shown on the last frame of drinking activity for visual purpose only.

Fig. 7: Detected C-STIP on drinking activity.

Recognition system finds RGB cuboid descriptors (RGB-DESC) centered at interest point (x, y, t). RGB cuboids descriptor is a 3 dimensional $(\Delta x, \Delta y, \Delta t)$. The sizes of cuboids Δx, Δy, and Δt, are defined by equations (15) and (16).

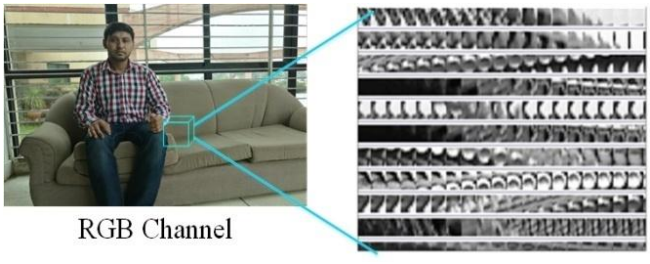

Fig. 8: RGB cuboids.

$$\Delta x = \Delta y = 2.\ ceil(3\sigma_c) + 1 \quad (15)$$

$$\Delta t = 2 \cdot ceil(3\tau_c) + \quad (16)$$

σ_c and τ_c are spatial and temporal scales at which interest point was detected.

Fig. 8 shows some of the detected cuboids. Every cuboid is flattened with respect to the temporal dimension; rows represent different cuboid and column represents its temporal location. Each Cuboid is normalized, and its Gradients are computed along x, y and z-direction. Three Gradient channels G_x, G_y, G_z will have the same dimension as the cuboid. Each cuboid pixel corresponding gradient values are concatenated to form a1D feature vector.

3.3 Combined RGBD Features

We tested human activity recognition algorithm using depth and RGB channels independently and by combining both channels information in which the depth interest points are mapped to RGB data. It has been concluded that the combined approach increases the recognition accuracy. Recent advancement in RGB depth sensor, like Kinect, provides both depth and RGB video concurrently without any orientation and time delay. This advantage enables researchers to extract interest points only from one channel and use it on multiple channels. Fig. 9 shows how drinking activity D-STIP's are mapped to RGB data.

Fig. 9: Mapped D-STIP to RGB channel.

When we extract interest, points using the only RGB channel, some false motion interest points can be seen on the background objects, as shown in Fig. 7. Consequently, less recognition accuracy is obtained while using the only RGB channel. By looking at MSR dataset, we assume that all the human activities are performed to a certain depth from the camera and one can get better interest points by just considering the foreground depth range. In the RGB channel, the background motion cannot be eliminated simply by considering the foreground region. While in depth channel, object motion in background can be eliminated by thresholding over a certain foreground depth range. Depth data provides better recognition accuracy due to an additional true motion interest points and for this reason we used D-STIP in a combined approach to increase the recognition accuracy. Our assumption can be proved by the fact that the recognition accuracy of combing RGB-STIP with RGB-DESC is worse than D-STIP and D-DESC combination.

RGBD feature combination is a three-step process:(i) The dimension of each channel descriptor is reduced using Principal Component Analysis (PCA) [38] to equalize the size, avoid the dimension curse and have an equal effect on the combined descriptor. (ii) Each channel's reduced feature vector is normalized to have an equivalent effect on RGB-Depth combination. (iii) The feature vector of RGB and Depth features from step 1 and 2 are concatenated. To reduce the dimension of both channel descriptors, we used PCA. Proposed research work is based on Ke and Suktankar [39] PCA-SIFT descriptors in which they reduced the dimension of SIFT from 128 to 12 without sacrificing the recognition accuracy. PCA is a statistical procedure that transforms correlated data to uncorrelated data, known as principal components. All the future computations are performed using 300 PCA value.

Reduced descriptors from RGB and depth data corresponding to the interest point, centered at (x, y, t) are normalized and concatenated. The proposed algorithm tests recognition accuracy on the combined RGB-depth features and individual channel features as well.

- In combine RGB-depth feature approach, interest points are only extracted from the depth data (D-STIP's). These points are used to acquire gradient based cuboid from the RGB data (RGB-DESC) and depth local histogram-based cuboid (D-DESC) from the depth data.
- In individual RGB channel, RGB-STIPs are extracted from the RGB channel using Gabor and Gaussian filters and are represented using RGB-DESC.
- In individual depth channel, D-STIPs are extracted from the depth channel and are represented using D-DESC.

3.4 Bag of the Visual Word (BOVW)

BOVW is a histogram vector corresponding to the feature vector occurrence in vocabulary. Salton and McGill [40] designed bag of words approach to distinguish words essay. The BOVW approach is applied to the reduced-concatenated RGB-Depth channel. It can also be applied to individual channels and combine them later, but it will be computationally inefficient. The BOVW approach has training and testing stages. The following steps explain the bag of words approach:

- *Step 1*, Training: Feature vectors from training data of multiple activities are computed using interest points and their descriptors.
- *Step 2*, Training: Feature vectors are clustered together to form a codebook. Codeword's are centers of these clustered data. The size of codebook plays an important role in the recognition accuracy; too small size will not represent all feature patches; too large feature size will produce quantization artifacts.
- *Step 3*, Testing: For a new video, an activity computes its descriptors.
- *Step 4*, Testing: Map each feature in the testing activity to visual the word in the codebook by computing minimum distance.
- *Step 5*, Testing: Compute histogram of words to find how often any word did appear.

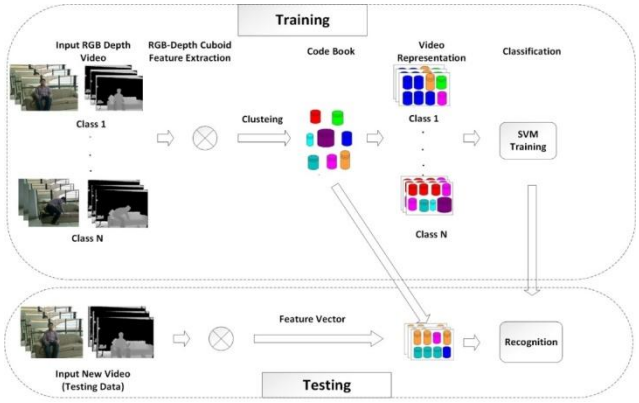

Fig. 10: Bag of visual words with SVM.

Bag of words approach represents each activity with its signature word, where each activity is represented by only 1D signature. To classify these activities, we used one vs. one multiclass linear support vector machine (SVM) [41] with cross-validation. Fig. 10 explains bags of visual words process with SVM to classify human activities in combined RGB-depth channels.

4. Results and Analysis

This section describes the experimental setup and results are obtained for the proposed RGBD human activity recognition algorithm. The algorithm is tested on the publically available MSR activity dataset [3], using Intel Processor Core i5 with MATLAB Tool. The results are compared with different low level and high-level activity recognition methods. Furthermore, different choices of parameters and their recognition accuracies are also computed. Experiments show that the combination of RGB depth features provides promising results.

4.1 MSR Activity Dataset

MSR activity dataset is collected to test the daily human activities in a more realistic setting, such that the background objects and human motion are also involved. The dataset is collected using Microsoft Kinect device in an indoor environment,and it provides both RGB and depth videos of each activity. Each activity is performed twice, once in sitting and once in standing position. The proposed activity recognition system is tested on 10 activities: drinking, eating, cheer up, calling, use a vacuum cleaner, lay down on the sofa, walking, sit down, stand up, toss the paper. Each subject performs sitting up and standing activities twice in different poses. There is a sofa in the background scene. Every activity is performed 10 times in sitting and 10 times standing position by different humans. Recognition system use 1-5 sequences for training and 6-10 sequences for testing purposes. Instead of having more training data, we have kept same percentage of training and testing sequences to make a rationale analysis of our approach.

Fig. 11: Detected D-STIP on MSR activities.

4.2 Evaluation Schemes

Cuboid features for depth feature extraction were proposed by Xia and Aggarwal [2] and they achieved very good results in MSR human activity recognition for depth channel only. For depth feature extraction, the proposed algorithm uses thesimilar parameter as explained in [2]. In MSR activity 3D dataset, RGBD activity recognition system provides the best accuracy of 92% when D-STIP's are extracted using $\sigma_d = 5{,}10$, $\tau_d = \frac{T}{17}$ and $Np = 500$, D-DESC are extracted using $n_{xy} = 4$ and $n_t = 3$, RGB-DESC are extracted with $\sigma_c = 2$ and $\tau_c = 4$. Correction function removes most of the noisy interest points from the depth data. Fig. 11 shows the detected D-STIP from each activity in the depth map data. Interest points are mapped to x, y domains and have been shown on the last frame in video. RGB and depth PCA are computed using 20,000 different patches from the respective channels. In each activity, sequences 1-5 are used for training purpose and 6-10 are used for testing purpose. Bag of visual words is computed by performing K-means clustering [42] on training data with bin size $k = 1200$ and we used LibSVM [43] one vs one approach to classify human activities.

The algorithm shows combined RGBD features provide significantly improved results as compared to single-channel approach. In combine RGBD approach, detected D-STIP's are mapped to RGB data, and D-DESC and RGB-DESC are computed around these points. Fig.12 shows confusion matrix for state of the art results on combine RGBD feature approach.

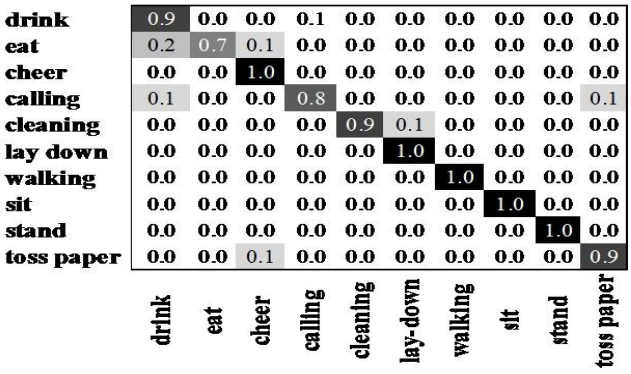

Fig. 12: Combined RGBD features confusion matrix.

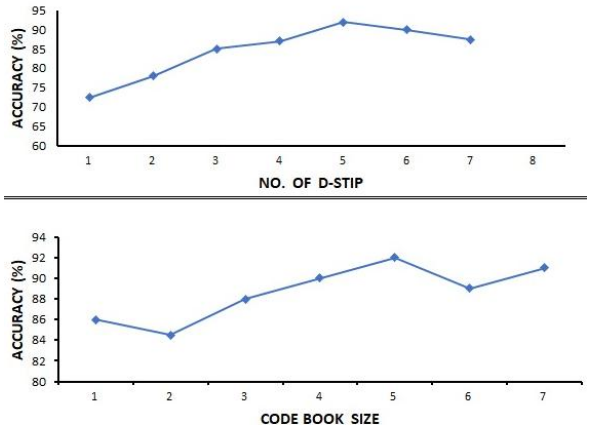

Fig. 13: Recognition results on different code book size and D-STIP values.

Fig. 13 shows the performance of combine RGBD human activity recognition system on MSR activity dataset with different parameter selection. Best results are computed using number of D-STIP, $Np = 500$ and code book size $k = 1200$.

Table 1: Single and combined approaches accuracy.

Approach		Accuracy%
Interest Point	Descriptor	
RGB	RGB	72
Depth	Depth	83
Depth	RGBD	92

Table 1 proves the capability of combined approach over a single channel approach, individually RGB channel provides accuracy of only 72% and depth channel provides accuracy of 83%. When both RGB and depth feature are combined, the proposed activity recognition system shows significant improvement in recognition accuracy and provides 92% accuracy.

Table 2 shows a comparison of different algorithm's recognition accuracy on MSR activity dataset. Unlike LOP [25] and HOND 4D [30], our system does not require any skeleton information or segmented human bodies.

Table 2: Recognition accuracy on MSR activity dataset.

Method	Accuracy%
LOP feature [26]	42.5
DTW [44]	54
Joint Position Feature [26]	68
HOG [13]	79.1
HOND 4D [31]	80
DCSF [2]	83
LOP+ Joint [26]	86
DCSF + Joints [2]	88.2
Combined RGBD (Ours)	92

Table 3: Computational cost comparison.

Descriptor	Time Cost (s) 500 points
RGB Cuboid (RGB-DESC)	16
Depth Cuboid(D-DESC)	60
Combined RGBD	76
SIFT 3D [16]	102

Combining Depth and RGB channel information certainly adds some computational cost. Table 3 shows the computational cost comparison with famous SIFT 3D [16] low-level descriptors. For a fair comparison, all descriptors are computed on the MATLAB platform with Intel Core i5 2.67, 3GB RAM, 32-bit system. SIFT 3D computational cost is only for RGB channel. It is notable that the combined approach is less computational expensive as compared to any state-of-the-art SIFT3D.

Table 2 shows that the proposed approach provides better results as compared to combining depth features with joints position. It can be noticed that the proposed algorithm is not dependent on skeletons or preprocessing methods like segmentation. Our algorithm provides a more generic approach when both RGB and depth information are accessible and synchronized. Furthermore, it can also be used with a wider range of human activities, e.g., group activity, gesture recognition or human-objects interaction activities.

5. Conclusion

In this research work, we compared the performance of combined RGBD features with a single channel-based feature in human activity recognition application. It has been observed that the best performance was achieved when the interest points are extracted only from the depth channel and identical points are used to extract the cuboid descriptors from both channels. Furthermore, the channel features are combined and a bag of visual words approach is used to recognize the human activity. The proposed framework is more realistic to recognize human activities in 4D (x, y, z, t) by using both channel information. It has been tested on the challenging MSR activity dataset and has provided better results in comparison with any state-of-the-art algorithm, based on low level or high-level features. The framework is more widely applicable on the color-depth senor based human activity recognition. A typical example is the surveillance system where the camera is mounted higher than the human body and both channels are synchronized.

Availability of less training data is a challenge in activity recognition task. Consequently, the future work is possible to make more such dataset. Furthermore, combining RGB depth features certainly add some computational cost. Therefore, one probable direction is to concatenate the RGB depth video or cuboid regions pixel values and then apply feature extraction on the combined video. Recognizing activities from compressed RGB-depth

video stream can be another alternative. The proposed approach can also be used when the skeleton joints information is obtainable. Joint points can be regarded as interest points and cuboids descriptors can be extracted from RGB and the depth videos around these points to recognize human activities. Proposed research work can be extended to recognize human activities by concatenating features from different channels, e.g., thermal infrared, depth and color sensor-based videos. The proposed approach can be experimented in variety of real-world scenarios, e.g., human action recognition for behaviour analysis, patient monitoring, and employee performance analysis and anomaly detection in intelligent video surveillance applications.

Acknowledgment

Authors are thankful to the Directorate of advanced studies and research for providing us the opportunity to carry out this research.

References

[1] R. Poppe, "A survey on vision-based human action recognition, "Image and vision computing", vol. 28, no. 6, pp. 976-990, 2010.

[2] Xia, Lu, and J.K. Aggarwal. "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

[3] L. Wanqing, Z. Zhang and Z. Liu, "Action recognition based on a bag of 3d points", IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2010.

[4] J.K. Aggarwal and M.S. Ryoo. "Human activity analysis: A review", ACM Computing Surveys (CSUR), vol. 43, no. 3, pp. 16, 2011.

[5] H. Chris and M. Stephens, "A combined corner and edge detector", Alvey Vision Conference, vol. 15, 1988.

[6] F. Robert, P. Perona and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning", Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, 2003.

[7] G.L. David, "Distinctive image features from scale-invariant key points, "Int. J. Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.

[8] B. Herbert, T. Tuytelaars and L. V. Gool, "Surf: speeded up robust features", Computer Vision–ECCV, Springer Berlin Heidelberg, pp. 404-417, 2006.

[9] L. Ivan, "On space-time interest points", Int. J. Computer Vision, vol. 64, nos. 2-3, pp. 107-123, 2005.

[10] D. Piotr, R. Vincent, G. Cottrell and B. Serge, "Behavior recognition via sparse spatio-temporal features", 2nd Joint IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005.

[11] G. Willems, T. Tuytelaars and L.V. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector", Computer Vision–ECCV, Springer Berlin Heidelberg, pp. 650-663, 2008.

[12] H. Wang, M.M. Ullah, A. Klaser, I. Laptev and C. Schmid, "Evaluation of local spatio-temporal features for action recognition", British Machine Vision Conference, 2009.

[13] I. Laptev, M. Marszelek, C. Schmid and B. Rozenfeld, "Learning realistic human actions from movies", IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[14] A. Klaeser, M. Marszalek and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients", Eds: M. Everingham and C. Needham, Proc. of the British Machine Conference, pp. 99.1-99.10, BMVA Press, September 2008.

[15] X. Jiang, T. Sun, B. Feng and C. Jiang, "A space-time surf descriptor and its application to action recognition with video words", IEEE 8th Int. Conf. on Fuzzy Systems and Knowledge Discovery, vol. 3, 2011.

[16] P. Scovanner, S. Ali and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition", Proc. of the 15th Int. Conf. on Multimedia, Augsburg, Germany, Sept. 23 - 28, 2007.

[17] W. LiMin, Y. Qiao and X. Tang, "Motionlets: Mid-level 3D parts for human motion recognition", IEEE Conf. on Computer Vision and Pattern Recognition, Portland, OR, pp. 2674-2681, 2013.

[18] Bobick, F. Aaron and J.W. Davis, "The recognition of human movement using temporal templates", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 3, pp. 257-267, 2001.

[19] K. Yan, R. Sukthankar and M. Hebert, "Spatio-temporal shape and flow correlation for action recognition", IEEE Conf. on Computer Vision and Pattern Recognition, Minneapolis, MN, pp. 1-8. 2007.

[20] S. Yaser, M. Sheikh and M. Shah, "Exploring the space of a human action", Tenth IEEE International Conference on Computer Vision, ICCV 2005, vol. 1, 2005.

[21] V. Ashok, R. Chellappa and A.K. Roy-Chowdhury, "The function space of an activity", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 959-968, pp. 117-22, 2006, New York, USA.

[22] M.H. Kolekar and D.P. Dash, "Hidden Markov Model based human activity recognition using shape and optical flow-based features, IEEE Region 10 Conference (TENCON), Singapore, pp. 393-397, 2016.

[23] Z. Cheng, L. Qin, Y. Ye, Q. Huang and Q. Tian, "Human daily action analysis with multi-view and color-depth data", Computer Vision–ECCV, Workshops and Demonstrations, Springer Berlin Heidelberg, 2012.

[24] A.A. Hadid and M. Pietikäinen, "Face recognition with local binary patterns", Computer Vision-eccv, Springer Berlin Heidelberg, pp. 469-481, 2004.

[25] Y. Zhao, L. Zicheng and C. Hong, "RGB-Depth feature for 3D human activity recognition", China Communications, vol. 10, pp. 93-103, 2013.

[26] J. Wang, L. Zicheng, J. Chorowski, C. Zhuoyuan and W. Ying, "Mining actionlet ensemble for action recognition with depth cameras, IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, pp. 1290-1297, 2012.

[27] J. Wang, L. Zicheng, J. Chorowski, Z. Chen and W. Ying, "Robust 3d action recognition with random occupancy patterns", Computer Vision–ECCV, Springer Berlin Heidelberg, 872-885, 2012.

[28] A.W. Vieira, E.R. Nascimento, G.L. Oliver, Z. Liu and M.F. Campos, "Stop: Space-time occupancy patterns for 3D action recognition from depth map sequences", Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Springer Berlin Heidelberg, pp. 252-259, 2012.

[29] Y. Xiaodong, C. Zhang and Y. Li Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients", Proc. of the 20th ACM International Conference on Multimedia. ACM, 2012.

[30] D. Navneet and B. Triggs, "Histograms of oriented gradients for human detection", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, 2005.

[31] O. Omar and Z. Liu, "Hon4d: Histogram of oriented 4D normal for activity recognition from depth sequences", IEEE Conference on Computer Vision and Pattern Recognition, 2013.

[32] L. Seidenari, V. Varano, S. Berretti and P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses", IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2013.

[33] M. Zanfir, M. Leordeanu and C. Sminchisescu, "The moving pose: an efficient 3D kinematics descriptor for low-latency action recognition and detection", IEEE International Conference on Computer Vision, pp. 2752-2759, 2013.

[34] J. Sung, C. Ponse, B. Suleman and A. Sexena, "Human activity detection from RGBD images", Plan, Activity and Intent Recognition, vol. 64, 2011.

[35] N. Bingbing, G. Wang and P. Moulin, "Rgbd-hudaact: A color-depth video database for human daily activity recognition", Consumer Depth Cameras for Computer Vision. Springer London, pp. 193-208, 2013.

[36] K.H. Swetha, R. Gupta and A. Saxena, "Learning human activities and object affordances from rgb-d videos", The Int. J. Robotics Res., vol. 32, no. 8, pp. 951-970, 2013.

[37] H.G. Gösta and H. Knutsson, "Signal processing for computer vision, vol. 2, 1995, Dordrecht: Kluwer Academic Publishers.

[38] F. Jerome, T. Hastie and R. Tibshirani, "The elements of statistical learning, New York: Springer Series in Statistics, vol. 1, 2001.

[39] K. Yan and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors", Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, 2004.

[40] S. Gerard and M.J. McGill, "Introduction to modern information retrieval", 1983.

[41] C. Corinna and V. Vapnik, "Support-vector networks", Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.

[42] J.A. Hartigan and M.A. Wong, "Algorithm AS 136: A k-means clustering algorithm", Applied Statistics, pp. 100-108, 1979.

[43] C. Chih-Chung and L. Chih-Jen, "LIBSVM: a library for support vector machines", ACM Transactions on Intelligent Systems and Technology, vol, 2, no. 3, pp. 27, 2011.

[44] M. Meinard and T. Röder, "Motion templates for automatic classification and retrieval of motion capture data", Proc. of the ACM SIGGRAPH/Euro Graphics Symposium on Computer Animation, Euro Graphics Association, 2006.