Towards Privacy-Preserving Activity Recognition Using Extremely Low Temporal and Spatial Resolution Cameras

Ji Dai, Jonathan Wu, Behrouz Saghafi, Janusz Konrad, Prakash Ishwar [*]
Department of Electrical and Computer Engineering, Boston University
8 Saint Mary's Street, Boston, MA, 02215
[jidai,jonwu,bsk,jkonrad,pi]@bu.edu

Abstract

Although extensive research on action recognition has been carried out using standard video cameras, little work has explored recognition performance at extremely low temporal or spatial camera resolutions. Reliable action recognition in such a "degraded" environment would promote the development of privacy-preserving smart rooms that would facilitate intelligent interaction with its occupants while mitigating privacy concerns. This paper aims to explore the trade-off between action recognition performance, number of cameras, and temporal and spatial resolution in a smart-room environment. As it is impractical to build a physical platform to test every combination of camera positions and resolutions, we use a graphics engine (Unity3D©) to simulate a room with various avatars animated using motions captured from real subjects with a Kinect v2 sensor. We study the performance impact of spatial resolutions from a single pixel up-to 10×10 pixels, the impact of temporal resolutions from 2 Hz up-to 30 Hz and the impact of using up-to 5 ceiling cameras. We found that reliable action recognition for smart-room centric gestures can still occur in environments with extremely low temporal and spatial resolutions. When using 5, single-pixel cameras at 30Hz we achieved a correct classification rate (CCR) of 75.70% across 9 actions, only 13.9% lower than the CCR for the same camera setup at 10×10 pixels. We also found that, in terms of the impact on action recognition performance, spatial resolution has the highest impact, followed by number of cameras, and temporal resolution (frame rate).

1. Introduction

Smart spaces of the future, whether at home or at work, promise to bring improved energy efficiency, health benefits, and productivity. For instance, energy savings can be realized by lowering illumination in regions void of humans, while health benefits can be realized by optimizing lighting conditions for specific activities, e.g., reducing screen glare when working on a laptop. Productivity can be improved, for example, by localizing occupants in order to maximize throughput rates in visible-light communication between wall- or ceiling-mounted transceivers and mobile devices. Furthermore, hand gestures can be used to control various room conditions like temperature, light, sound, etc.

In addition to the detection, localization, and tracking of human subjects, reliable *recognition* of human activities and gestures is crucial for realizing the full spectrum of the aforementioned benefits. Human activity and gesture recognition has been extensively studied in the computer vision and signal processing communities. However, most of the work is based on video cameras where there is little to no expectation of occupant privacy.

One way to address privacy concerns is to degrade (reduce) the quantity and quality of the data gathered to the point where it no longer provides any "visual utility" to eavesdroppers. However, this will also degrade action recognition accuracy. The number of cameras and their spatial and temporal (frame rate) resolutions are three important data dimensions that significantly impact both "visual utility" and action recognition accuracy. What are the limits to which we can reduce these data dimensions without significantly impacting action recognition accuracy? What are the ranges of tradeoff between performance and resolution along different data dimensions? To which data dimension(s) is the performance most sensitive?

In this paper, we consider scenarios with 1 to 5 grayscale cameras with no more than 10×10 spatial resolution and $2-30$ Hz frame rate. Is there any hope of getting any useful action recognition performance in such extreme scenarios? We explore and attempt to answer these questions in the remainder of this paper.

[*] This work was supported by the NSF under Smart Lighting ERC Cooperative Agreement No. EEC-0812056.

2. Related Work

Over the last decade or so, a vast amount of literature has been published on action recognition from video (we refer the interested reader to the survey by Aggarwal and Ryoo [4]). Most of these publications have dealt with extracting discriminative features from high spatio-temporal resolution video. These features have most commonly been based on optical flow [9], point trajectories [6], silhouettes [11, 15, 18], and spatio-temporal interest points [8, 14].

There are, however, a few veins of action recognition research that have focused on video with low spatial or low temporal resolution (but not both simultaneously). In terms of low spatial resolution, Efros *et al.* [9] applied optical flow with frame-to-frame normalized correlation in time on subjects as small as 30 pixels in height. Similarly, Ahad *et al.* [5] used directional motion history images on subjects around 24 pixels in height. However, both these techniques are ill-suited at an *even lower* resolution which we call *extremely low resolution*. At extremely low spatial resolutions of 10×10 (with subjects as small as 7 pixels in height) down to 1×1 (a single pixel), optical flow cannot be reliably (or even meaningfully) computed.

Similarly, it has been shown that action recognition performance can be reliable at low frame rates, even when only a few key frames can be found (in extreme cases, even a single still frame) [7, 10, 16]. Although these works do not directly and systematically quantify the impact of having a very low frame rate on recognition performance, their results are one of the motivating factors for our own investigation. The most related work in this context is by Harjanto *et al.* [12], who conclude that the impact of frame rate varies depending on feature selection and the environment and actions in the dataset. From the perspective of our work, however, their study is limited to actions with standard spatial resolutions as they did not study the effect of low frame rate *combined with* low spatial resolution.

With regard to privacy preservation through the use of low-resolution data, perhaps the most closely related work is that of Jia and Radke [13], who explore privacy-preserving tracking and coarse pose estimation using a network of ceiling-mounted *time-of-flight* sensors. Similarly, Tao *et al.* [17] use a network of ceiling-mounted binary *passive infrared sensors* to recognize a set of daily activities. However, activities in their dataset are only performed in known fixed areas of the room. This activity-location dependence is effectively utilized in their approach but may not be realistic in practice.

3. Overview of Methodology

There exist overwhelmingly many camera configurations that can affect action recognition performance. For example, one could characterize a camera configuration by the number of cameras, their resolution, frame rate, shutter speed, aperture, zoom level, lens distortion, positions and orientations relative to each other and the room, etc.; the list goes on. In this paper, we study only a small but important set of these configurations. Specifically, we study the impact of the number of cameras and their spatial and temporal resolutions on action recognition performance.

Since it is impractical to test every configuration in a real testbed, we used a graphically-rendered smart room in Unity3D© [3] to animate avatars. Fig. 1 shows our simulated smart room from one viewpoint. However, for the avatar animations, we used true human motions captured from subjects facing a Kinect v2 [2] sensor. The animated avatars were captured from 5 viewpoints with varying spatial and temporal resolutions and the resulting videos were used to evaluate activity recognition performance.

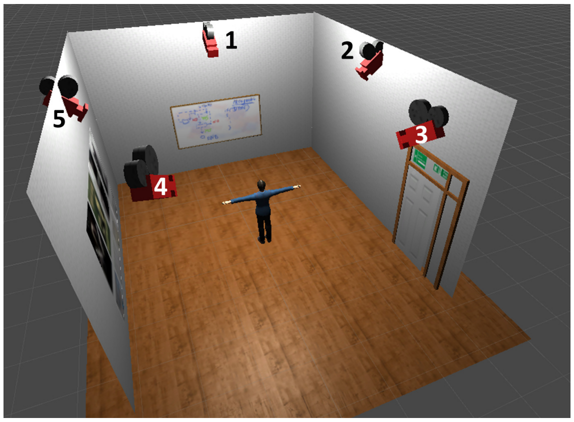

Figure 1. A seminar room simulated in Unity3D© with 5 ceiling-mounted cameras in a pentagonal arrangement (cameras are numbered accordingly). A single omni-directional light source was used to illuminate the entire room. For visualization, this light source has higher intensity than shown in Figs. 2 and 3.

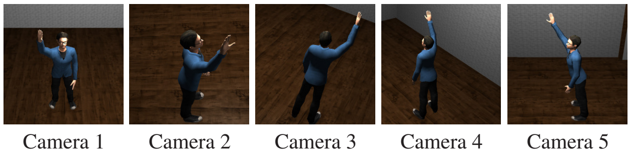

Figure 2. All camera viewpoints of an avatar raising his arm. These camera viewpoints correspond to the ones shown in Fig. 1.

In this study, we assume that there is a single avatar in the room and the zoom setting of each camera is such that the avatar encompasses much of the field of view (see Fig. 2). We also assume that all training and test avatars are roughly at the same position and orientation relative to the camera network. Finally, we assume that all cameras have the same (common) spatial and temporal resolutions. With this setup, we vary the number of cameras (from a pentagonal arrangement of 5, to a subset of 3 out of the five, down to 1 out of the five), their spatial resolution (from 10×10 down to 1×1), and their temporal resolution (from 30 Hz to 2 Hz, including different time offsets) and evaluate the action recognition accuracy.

3.1. Action Recognition Algorithm

In order to evaluate action recognition performance at such extremely low resolutions, we developed a simple pixel-wise, time-series based algorithm. The use of pixel-wise time series is motivated by the lack of reliable estimation algorithms for common features such as optical flow or spatio-temporal interest points at the extremely low resolutions that we study.

At a high level, our approach is based on extracting, from a grayscale video sequence, a spatially-aligned and grayscale-normalized rectangular spatio-temporal video cuboid which tightly encompasses the avatar's silhouette tunnel (i.e., a sequence of silhouettes [11]). This accounts for some of the global spatial misalignments between different action instances, and reduces inter-avatar appearance variability (e.g., due to clothing).

In more detail, for each grayscale video sequence, we first extract a sequence of binary silhouettes by thresholding the absolute deviation of the grayscale value at each pixel from the known background value (which is known due to our Unity3D© simulation but in general can be estimated). Next, the smallest-volume rectangular spatio-temporal video cuboid that contains the entirety of the background-subtracted silhouettes is found.[1] Effectively, this isolates a fixed spatial region of interest (ROI) across time by removing uninformative background pixels. Further, as the dimensions of the rectangular cuboids from different action sequences can be different, we resize all cuboids to a fixed spatial resolution of $R \times R$, with $R = 10$, and a fixed temporal length of $T = 500$. Spatial resizing is done through bi-cubic interpolation and temporal resizing is performed using cubic-spline interpolation.

Finally, a variant of mean-variance normalization is applied to each resized cuboid $\{x_{ijk}[t], i, j = 1, \ldots, R, t = 1, \ldots, T\}$, where $x_{ijk}[t]$ denotes the grayscale value of pixel (i, j) at time t in camera k, as follows:

$$\hat{x}_{ijk}[t] = \frac{x_{ijk}[t] - \mu_{ijk}}{\sigma_k}. \tag{1}$$

Here, μ_{ijk} denotes the "static" pixel mean value across time for the spatial location (i, j), and σ_k denotes the empirical standard deviation taken across *all* pixels in the cuboid for camera k. The subtraction of the mean emphasizes a subject's local dynamics in the cuboid (by removing an intensity bias), while the division by standard deviation partially accounts for the variability in subject's clothing.

Using these normalized cuboids, we propose a simplistic nearest-neighbor classifier using the l_1 distance to discriminate between action sequences. The distance between two normalized cuboids $\hat{x}$ and $\hat{y}$ is defined as:

$$d(\hat{x}, \hat{y}) = \sum_{k=1}^{K} \sum_{t=1}^{T} \sum_{i=1}^{R} \sum_{j=1}^{R} |\hat{x}_{ijk}[t] - \hat{y}_{ijk}[t]|, \tag{2}$$

where the distance is taken across all K cameras, each with its own unique cuboid of spatial dimensions $R \times R$ and temporal dimension T (chosen as 10×10 and 500 respectively).

4. Dataset and Simulation

In order to render realistic human motion in a virtual smart-room environment, a small dataset of actions typical of seminar-room scenario were collected. These action samples were collected from 12 subjects (7 males and 5 females, mostly college-aged), each of whom performed 9 distinct actions, repeating each action 3 times. All action samples were performed while standing at a rest position (hands at the side) in front of a single forward-facing Kinect v2 depth camera. Further, subjects were requested to leave and re-enter the field of view of the camera so as to reduce position and pose biases between samples.

The 9 seminar-room centric actions that were collected are: *Answering Mobile* (removing mobile from a pocket, unlocking it and raising it to ear level; ≈ 6-17 seconds), *Checking Mobile* (removing mobile from a pocket, unlocking it and manipulating the device to read email, surf the web, etc.; ≈ 10-17 seconds), *Raising Hand* (raising right hand straight up as if voting or asking a question; ≈ 3-7 seconds), *Lowering Hands* (quickly raising both hands up in tandem and then slowly lowering them in tandem as if requesting an audience to settle down; ≈ 3-9 seconds), *Raising Volume* (raising right hand up with the palm facing up as if requesting to increase the sound volume; ≈ 2-5 seconds), *Writing on Board* (≈ 4-12 seconds), *Clapping* (≈ 4-10 seconds), *Walking* (≈ 3-5 seconds), and *Sitting* (≈ 6-11 seconds).

Example frames with forward-facing views for each of these 9 actions are shown in Fig. 3. We note that some of these actions have segments where they look very similar, e.g., *Raising Hand* and *Raising Volume*. These were included intentionally in order to discover the extent to which similar-looking actions get confused at extremely low resolutions.

Once a set of realistic action samples were recorded, a virtual seminar-room environment was built in Unity3D© (see Fig. 1). The virtual environment simulated an avatar moving around the center of a room whose actions were captured by 5 grayscale ceiling-mounted cameras (arranged

[1] For image sequences with resolutions smaller than 3×3, this cuboid is typically the original sequence in its entirety (the silhouettes span the entire field of view).

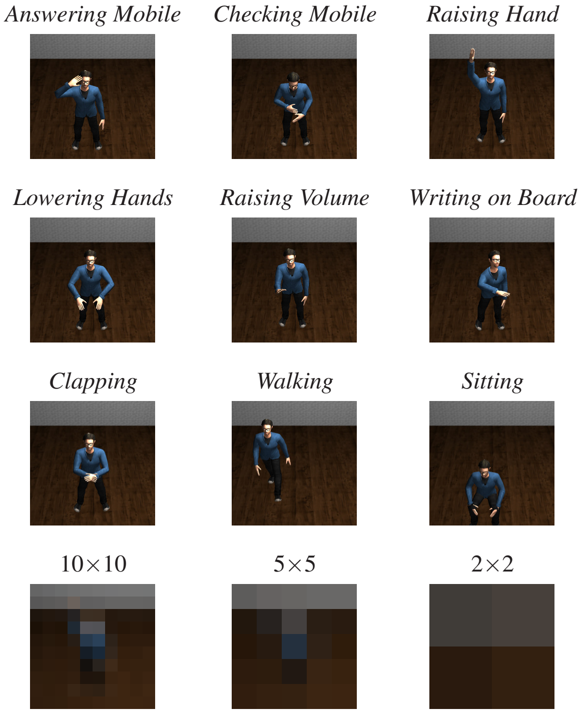

Figure 3. Discriminating poses of the 9 seminar-room centric actions from the forward-facing virtual camera (camera 1). The last row shows the extremely low spatial resolution versions of the *Raising Hand* pose.

in the shape of a pentagon pointing towards the avatar). Each camera was set up such that the avatar would be contained within its field of view (see Fig. 2). Various camera configurations of the ceiling-mounted cameras were tested: the number of cameras, and the temporal and spatial resolutions (fixed across cameras) were varied. The parameters and their values for these camera configurations are listed in Table 1. Notably, the avatar was positioned and animated based on the skeletal information of real-life subjects (from median filtered pose estimates obtained from the Kinect 2.0 SDK). Each sample of each subject was mapped through skeletal re-targeting onto multiple gender-matching avatars (shown in Fig. 4). Gender-matching was done because the skeletal proportions and avatar-builds of different genders are quite distinct. For evaluation, we only used a subset of these avatar mappings, which we describe in more detail in the following section.

Camera Configuration	Values
Spatial Resolution [pixels]	1×1,2×2,3×3,4×4,5×5,10×10
Temporal Resolution [Hz]	2, 7.5, 15, 30
Number of Cameras	1, 3, 5

Table 1. List of parameter values for camera configurations tested.

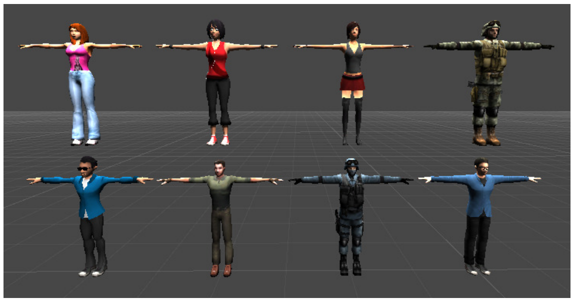

Figure 4. Various avatars (3 female, 5 male) used in Unity3D$^{©}$.

5. Experimental Evaluation

We report the average correct classification rate (CCR) for various camera configurations using a variant of leave-person-out cross-validation. This entails computing a running CCR across M subject-avatar matching iterations. In each iteration, each of the 12 subjects is sequentially selected to be the single "person-out" test subject. Then, one same-gender avatar is randomly assigned to the test subject and the assigned avatar and all 27 action samples from that test subject (9 actions × 3 repetitions) are removed from the pool. Next, the remaining training subjects are randomly assigned same-gender avatars (different from test-subject's avatar) by sampling with replacement. Then, each of the 27 samples of the test avatar is classified using the 297 training samples (27 samples × 11 training subjects) which yields an average CCR for this test subject. By averaging the CCR score across all 12 test subjects we get an average CCR for one iteration. This process is repeated for $M = 100$ iterations in order to obtain the final average CCR. We perform this testing process for all the camera configurations that we evaluate.

Notably, this procedure evaluates subject motion across a range of avatars, each varying in build and shape. Furthermore, this "leave-person-out" procedure is powerful in showing the robustness of our results, as the classifier never sees samples from the matched avatar *and* subject used for testing.

5.1. Impact of Spatial Resolution

The impact of spatial resolution change was evaluated for the resolutions ranging from 10×10 down to 1×1 across 5 cameras. The results of our method along with action-specific CCRs are shown in Table 2.

Fig. 5 shows the mean-variance normalized time series (see equation (1)) for all 9 actions at single-pixel resolution. These plots highlight the differences that occur even when the same action and avatar are used but different subjects drive the same avatar's motions. The intensity signals are similar for the same action yet exhibit non-negligible variations thus further confirming the richness of our dataset.

From the highest resolution 10×10 down to 1×1, there

Spatial Resolution	**CCR**	StdDev	Answering Mobile	Checking Mobile	Raising Hand	Lowering Hands	Raising Volume	Writing on Board	Clapping	Walking	Sitting
10×10	**89.60%**	1.53%	66.92%	80.19%	97.56%	99.61%	95.72%	85.25%	83.08%	99.83%	98.28%
5×5	**86.80%**	1.61%	68.11%	74.69%	95.31%	100.00%	90.17%	79.25%	75.39%	99.72%	98.58%
4×4	**84.52%**	1.58%	68.42%	62.64%	88.03%	100.00%	91.44%	78.89%	74.03%	99.64%	97.56%
3×3	**86.71%**	1.94%	77.22%	69.64%	91.11%	99.36%	89.33%	80.00%	76.56%	99.86%	97.28%
2×2	**82.71%**	1.96%	70.58%	45.64%	87.94%	99.75%	88.06%	76.56%	82.11%	99.64%	94.08%
1×1	**75.70%**	1.88%	50.92%	35.42%	85.53%	98.36%	85.69%	58.56%	79.81%	94.08%	92.94%

Table 2. Average CCR for various spatial resolutions $R \times R$ ranging from 10×10 to 1×1. The temporal resolution is fixed at 30 Hz and the number of cameras is fixed to be 5. We have also shown the average CCR for each of the 9 actions per configuration.

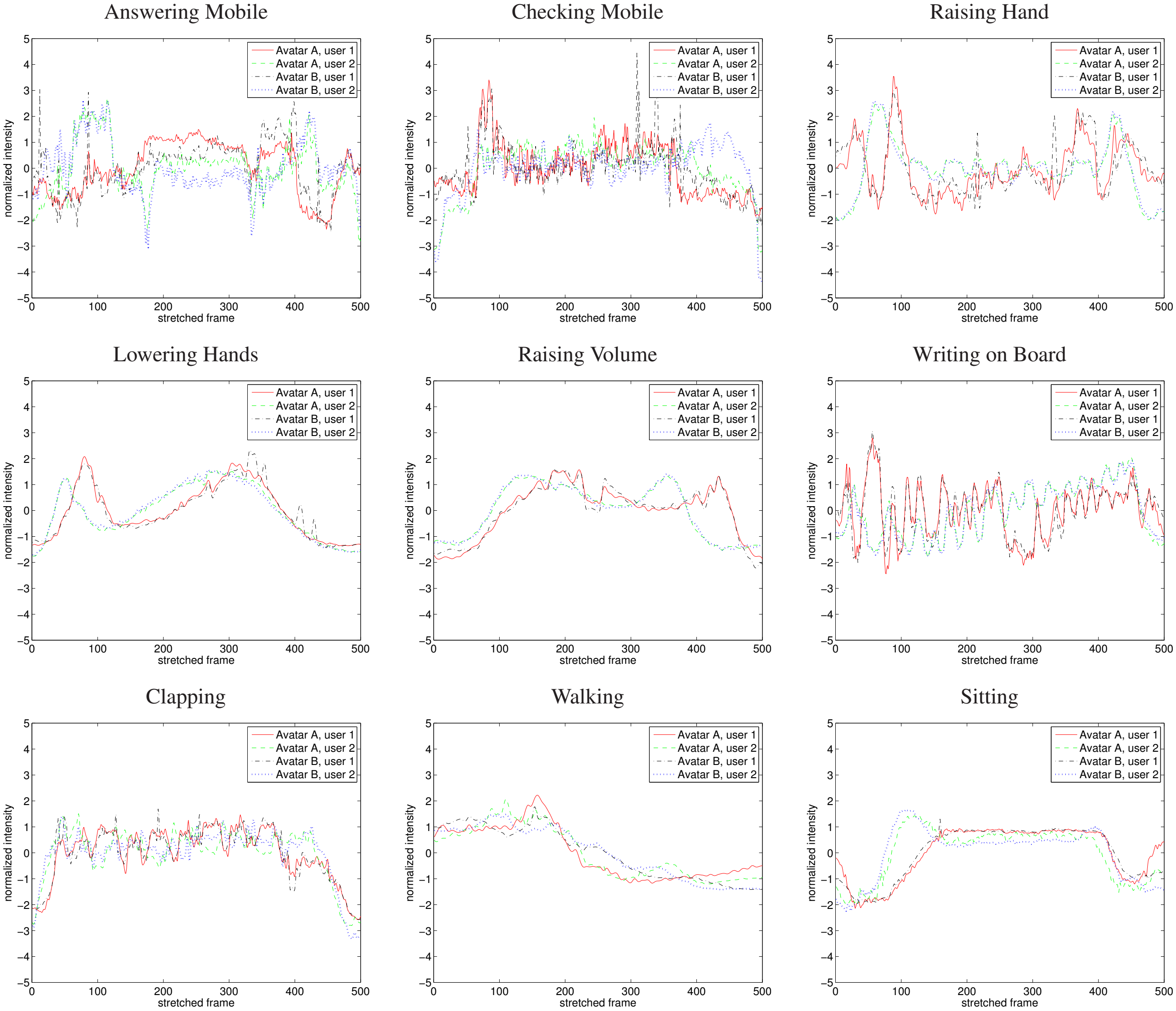

Figure 5. Normalized single-pixel grayscale intensities (see equation (1)) for all 9 actions from a fixed camera viewpoint. Shown are the intensities of 2 subjects mapped to the same 2 avatars.

is generally a decrease in performance due to the drop in spatial resolution. Although there is a slight performance increase at 3×3, we believe this is due to frequent imperfect alignments which our cuboid-based technique cannot perfectly compensate for.

In terms of individual actions, *Raising Hand*, *Lowering*

Hands, *Raising Volume*, *Walking*, and *Sitting* achieve consistent CCRs in the range of 85-100% across all tested resolutions. At the same time, *Answering Mobile* is the hardest action to recognize as its CCR ranges from 50% to 67%, primarily due to confusion with the similar action *Checking Mobile*. *Checking Mobile* and *Writing on Board* have quite high CCR at higher resolutions tested, but drop off dramatically at single-pixel resolution, which is likely due to intricate hand movements involved in both actions. Finally, *Clapping*, has a consistent CCR in the 70-80% range but quite a bit lower than that for the most reliable actions. This is due to the relatively small amplitude of hand movements extended in front of the body that can be partially confused with *Raising Hand*, *Lowering Hands*, and *Raising Volume*.

Overall, however, these results are promising. Even at extremely low spatial resolutions, good action recognition performance appears to be attainable. Even for the largest drop in spatial resolution from 10×10 to 1×1, there is a relatively small 13.9% reduction in performance.

5.2. Impact of Temporal Resolution

The impact of temporal resolution was evaluated for frame rates ranging from 30 Hz down to 2 Hz. A lower frame rate was achieved by down-sampling the original 30 Hz signal by a fixed integer constant. To account for potential bias due to temporal offsets, the performance from every possible temporal offset was evaluated and averaged together. The results along with action-specific classification accuracies are shown in Table 3.

The drop in performance due to the decrease in temporal resolution is quite small. Even for the largest drop in resolution, from 30 Hz down to 2 Hz, there is only a small 3.11% reduction in performance. This is not entirely surprising, as most of the seminar-room centric actions were arm- and body-based, so that large movements could still be picked up by a camera with a low frame rate (for example, some gestures had durations of up to 10 seconds). If, however, one had to distinguish between very intricate and quick hand motions, then this reduction of frame rate can be expected to produce a significant performance degradation.

Similarly to the spatial resolution change, the same 5 actions: *Raising Hand*, *Lowering Hands*, *Raising Volume*, *Walking*, and *Sitting* can be reliably recognized.

5.3. Impact of Camera Count and Arrangement

The impact of the number of cameras used and their arrangement was evaluated with the results reported in Table 4 along with action-specific classification accuracies in Table 5. For a single camera and for 3 cameras, all combinations of camera selections were tested (5 positions for a single camera, and "5 choose 3", i.e., 10 combinations for 3 cameras), with the resulting CCRs averaged and reported in our table. Not surprisingly, the best-performing single-camera viewpoint came from the frontal-facing camera number 1. In fact, all the top-performing 3-camera combinations included the frontal camera. In terms of notable trends due to the reduction of camera counts, there is a slight drop on average from 5 to 3 cameras (2.65% decrease), and a larger drop on average from 5 to 1 camera (11.64% decrease). As before, the same 5 actions can be reliably recognized when changing the number of cameras and their arrangement.

Camera Combination	CCR	StDev
1 (front)	**88.57%**	**1.89%**
2 (right)	71.29%	2.37%
3 (back right)	69.65%	1.92%
4 (back left)	78.08%	2.71%
5 (left)	82.21%	2.26%
Average for 1 camera	**77.96%**	**2.23%**
1-2-3	85.49%	1.88%
2-3-4	83.29%	1.58%
3-4-5	84.02%	1.83%
4-5-1	89.92%	1.87%
5-1-2	89.99%	1.67%
1-3-4	87.12%	1.61%
2-4-5	86.24%	1.91%
3-5-1	**90.15%**	**1.64%**
4-1-2	89.30%	1.88%
5-2-3	83.94%	1.63%
Average for 3 cameras	**86.95%**	**1.75%**
5 cameras	**89.60%**	**1.53%**

Table 4. Average CCR for various camera counts and combinations ranging from 5 to 1. The spatial resolution is fixed at 10×10 and the temporal resolution is fixed at 30 Hz. Notably, the top-performing camera combinations always contain the frontal camera (camera 1).

5.4. Summary of Results

Given the overwhelming amount of data provided in the earlier tables, we overview our results for the extreme camera configurations in Table 6 and their action-specific classification accuracies in Table 7.

There are a few notable trends that exist in every camera configuration we tested. Firstly, there exist a set of actions that consistently have reliable performance (normally above 85%, sometimes 100%). These actions are: *Raising Hand*, *Lowering Hands*, *Raising Volume*, *Walking*, and *Sitting*. On the other hand, *Answering Mobile* is the hardest action to recognize. In fact, *Answering Mobile* is frequently confused with the similar action *Checking Mobile*. Not surprisingly, these two actions suffer the most from a 2×2 spatial resolution reduction to 1×1, as does *Writing on Board* which involves intricate hand movements.

Temporal Resolution	CCR	StdDev	Answering Mobile	Checking Mobile	Raising Hand	Lowering Hands	Raising Volume	Writing on Board	Clapping	Walking	Sitting
30 Hz	**89.60%**	1.53%	66.92%	80.19%	97.56%	99.61%	95.72%	85.25%	83.08%	99.83%	98.28%
15 Hz	**89.35%**	1.44%	66.65%	80.26%	97.57%	99.61%	95.68%	84.42%	81.64%	99.88%	98.40%
7.5 Hz	**89.35%**	1.52%	67.77%	81.12%	97.29%	99.57%	96.40%	82.96%	80.69%	99.85%	98.52%
2 Hz	**86.49%**	1.79%	62.10%	77.47%	98.54%	98.68%	95.71%	74.95%	77.67%	97.34%	98.54%

Table 3. Average CCR for various temporal resolutions ranging from 30 Hz to 2 Hz through downsampling. The spatial resolution is fixed at 10×10 and the number of cameras is fixed to be 5. We have also shown the average CCR for each of the 9 actions per configuration.

Camera Combination	CCR	StdDev	Answering Mobile	Checking Mobile	Raising Hand	Lowering Hands	Raising Volume	Writing on Board	Clapping	Walking	Sitting
5	**89.60%**	1.53%	66.92%	80.19%	97.56%	99.61%	95.72%	85.25%	83.08%	99.83%	98.28%
Avg. of 3	**86.95%**	1.75%	59.61%	74.52%	95.83%	99.15%	95.39%	80.71%	80.37%	99.15%	97.40%
Avg. of 1	**77.96%**	2.23%	48.45%	61.05%	85.31%	94.58%	90.10%	70.16%	67.95%	95.33%	88.49%

Table 5. Average CCR for various camera counts and combinations ranging from 5 to 1. The spatial resolution is fixed at 10 × 10 and the temporal resolution is fixed at 30 Hz. We have also shown the average CCR for each of the 9 actions per configuration.

Description	Camera Configuration	CCR	StDev
Best	10×10, 30 Hz, 5 cams	89.60%	1.53%
Low Frame-rate	10×10, 2 Hz, 5 cams	86.49%	1.79%
Single Camera	10×10, 30 Hz, 1 cam	77.96%	2.23%
Low Spatial Res.	1×1, 30 Hz, 5 cams	75.70%	1.88%
Everything Low	1×1, 2 Hz, 1 cam	48.39%	2.34%

Table 6. Overview of CCRs for various camera configurations.

Although our results, in general, show a decrease in performance due to the drop in resolution and the number of cameras, we occasionally observe a slight departure from this monotonicity. We believe this is due to misalignments of avatar-activity areas between the frames under comparison which our coarse cuboid-based alignment technique (Section 3.1) cannot perfectly compensate.

6. Validation on Real Data

In order to validate our findings on real data, we repeated a subset of our experiments on the multi-view action sequences from the IXMAS action dataset [19] (5 cameras, 10 subjects, 12 actions). We used the static ROIs (64×48 pixels) provided by the dataset, where the subjects occupy most of the field of view, and decimated the spatial and temporal resolutions to simulate extremely low-resolution camera views (Fig. 6).

For experimental validation, we used the method proposed in Section 3.1, but with a different cuboid extraction technique. This is due to the fact that the method proposed earlier requires a known background which is not provided for the ROI sequences in IXMAS. For each video sequence, we extract the cuboid as follows. In order to determine the column boundaries of the cuboid, we select a threshold and find the first and last columns whose average temporal variance exceeds the threshold. By average temporal variance of a column we mean the sum of the variances (across time) of each pixel in the column divided by the total number of pixels in the column (i.e., total number of rows). Similarly, we detrmine the row boundaries of the cuboid as the first and last rows whose average temporal variance exceeds the (same) threshold. Effectively, this method computes a cuboid that only contains areas with significant motion. For example, if a subject's action sample does not have significant lower-body motion, then the cuboid would exclude that subject's lower body.

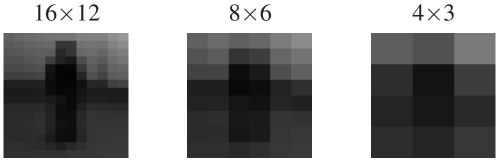

Figure 6. Extremely low spatial resolution frames generated from the IXMAS dataset (decimated versions of the provided ROIs from the frontal facing camera).

We evaluated action recognition performance for spatial resolutions of 16×12, 8×6, 4×3, and 1×1, temporal resolutions of 25 Hz and approximately 2 Hz (each 25 Hz sequence is down-sampled by 12 in time), and camera counts of 5 and 1 (frontal only). CCR was computed in each case using leave-person-out cross validation. The overview of our results are shown in Table 8.

These results on real data are qualitatively consistent with the results in Table 6 on synthetic data: the recognition rates improve with spatial and temporal resolutions and the number of cameras, and the CCR is quite insensitive to the decrease of temporal resolution with around a 5% CCR loss from 25 Hz to 2 Hz. The best-performing configuration has a CCR of 80%. While this is below the best CCR for synthetic data, it is quite comparable. The close qualitative

Camera	Answering	Checking	Raising	Lowering	Raising	Writing on	Clapping	Walking	Sitting
Configuration	Mobile	Mobile	Hand	Hands	Volume	Board			
10×10, 30 Hz, 5 cams	66.92%	80.19%	97.56%	99.61%	95.72%	85.25%	83.08%	99.83%	98.28%
10×10, 2 Hz, 5 cams	62.10%	77.47%	95.97%	98.68%	95.71%	74.95%	77.67%	97.34%	98.54%
10×10, 30 Hz, 1 cam	48.45%	61.05%	85.31%	94.58%	90.10%	70.16%	67.95%	95.33%	88.49%
1×1, 30 Hz, 5 cams	50.92%	35.42%	85.53%	98.36%	85.69%	58.56%	79.81%	94.08%	92.94%
1×1, 2 Hz, 1 cam	27.86%	20.45%	55.06%	79.18%	58.93%	27.71%	44.46%	65.32%	62.31%

Table 7. Average CCR for each of the 9 actions for various camera configurations.

Description	Camera Configuration	CCR	StDev
Best	16×12, 25 Hz, 5 cams	80.00%	6.90%
-	8×6, 25 Hz, 5 cams	77.78%	7.52%
-	4×3, 25 Hz, 5 cams	76.94%	8.39%
Low Frame-Rate	16×12, 2 Hz, 5 cams	74.35%	6.68%
Single Camera	16×12, 25 Hz, 1 cam	67.11%	7.91%
Low Spatial Res.	1×1, 25 Hz, 5 cams	63.33%	11.40%
Everything Low	1×1, 2 Hz, 1 cam	29.21%	2.61%

Table 8. Overview of CCRs for various camera configurations on the IXMAS ROI dataset.

agreement between the results on synthetic and real data is quite encouraging because it validates the simulation-based approach and provides evidence to support the conclusion that a simulation-based study can serve as a fairly reliable proxy for real-world data.

7. Conclusions

This paper investigated and empirically reported the impact of extremely low temporal and spatial resolution, the number of cameras used as well as their arrangement in relation to action recognition performance within a seminar-room scenario. Each factor was individually studied using the proposed algorithm, when the other two factors were fixed. From our study, we found that a reduction in spatial resolution seems to have the most drastic effect on recognition performance, followed by camera count/arrangement, and then temporal resolution. Further, we investigated the worst-case camera configuration we could: 1 camera, with a 1×1 spatial resolution, and a 2 Hz temporal resolution, and found a very low average CCR of 48.39% ± 2.34%. This result can be compared to our best performing camera configuration: 5 cameras, each with a 10×10 spatial resolution, and a 30 Hz temporal resolution, which had a CCR of 89.6% ± 1.53%. The difference in performance suggests that having a single camera, with almost no spatial or temporal resolution whatsoever is ill suited for action recognition. However, our work has shown that with a slight increase in resolution and a few additional low resolution cameras, a very reasonable recognition rate of around 80% can be achieved (see camera configuration: 5 cameras, 2×2, 30 Hz). Clearly, at this spatial resolution any privacy concerns are mitigated. Further, these results have also been validated on real data.

As a final point, this paper focused on understanding the trade-offs between temporal and spatial resolution and camera counts rather than seeking to find an algorithm that would yield the best CCR. Yet, even with a simplistic algorithm, we have achieved surprisingly good performance with what would typically be considered extremely low and unusable resolutions.

More information on this research as well as some resources are available at [1].

References

[1] Boston University: Privacy-Preserving Smart-Room Analytics. vip.bu.edu/projects/vsns/privacy-smartroom/, 2015. 8

[2] KinectSDK. www.microsoft.com/en-us/kinectforwindows/, 2015. 2

[3] Unity - Game Engine. unity3d.com, 2015. 2

[4] J. Aggarwal and M. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43(3):16:1–16:43, 2011. 2

[5] M. Ahad, J. K. Tan, H. S. Kim, and S. Ishikawa. A simple approach for low-resolution activity recognition. *International Journal for Computational Vision and Biomechanics*, 3(1):17–24, 2010. 2

[6] S. Ali, A. Basharat, and M. Shah. Chaotic invariants for human action recognition. In *International Conference on Computer Vision (ICCV)*, 2007. 2

[7] J. W. Davis and A. Tyagi. Minimal-latency human action recognition using reliable-inference. *Image and Vision Computing*, 24(5):455–472, 2006. 2

[8] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, 2005. 2

[9] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *International Conference on Computer Vision (ICCV)*, 2003. 2

[10] G. Guo and A. Lai. A survey on still image based human action recognition. *Pattern Recognition*, 47(10):3343–3361, 2014. 2

[11] K. Guo, P. Ishwar, and J. Konrad. Action recognition from video using feature covariance matrices. *IEEE Transactions on Image Processing*, 22(6):2479–2494, 2013. 2, 3

[12] F. Harjanto, Z. Wang, S. Lu, and D. D. Feng. Evaluating the impact of frame rate on video based human action recognition. In *ACM Conference on Image and Vision Computing New Zealand*, 2012. 2

[13] L. Jia and R. J. Radke. Using time-of-flight measurements for privacy-preserving tracking in a smart room. *Industrial Informatics, IEEE Transactions on*, 10(1):689–696, 2014. 2

[14] B. Saghafi, E. Farahzadeh, D. Rajan, and A. Sluzek. Embedding visual words into concept space for action and scene recognition. In *British Machine Vision Conference (BMVC)*, 2010. 2

[15] B. Saghafi and D. Rajan. Human action recognition using pose-based discriminant embedding. *Signal Processing: Image Communication*, 27(1):96–111, 2012. 2

[16] K. Schindler and L. Van Gool. Action snippets: How many frames does human action recognition require? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. 2

[17] S. Tao, M. Kudo, and H. Nonaka. Privacy-preserved behavior analysis and fall detection by an infrared ceiling sensor network. *Sensors*, 12(12):16920–16936, 2012. 2

[18] L. Wang and D. Suter. Learning and matching of dynamic shape manifolds for human action recognition. *IEEE Transactions on Image Processing*, 16(6):1646–1661, 2007. 2

[19] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2-3):249–257, 2006. 7