TOWARDS PRIVACY-PRESERVING RECOGNITION OF HUMAN ACTIVITIES

Ji Dai, Behrouz Saghafi, Jonathan Wu, Janusz Konrad, Prakash Ishwar *

Department of Electrical and Computer Engineering, Boston University
8 Saint Mary's Street, Boston, MA, 02215
[jidai,bsk,jonwu,jkonrad,pi]@bu.edu

ABSTRACT

A smart room of the future is expected to facilitate intelligent interaction with its occupants while respecting their privacy. Although standard video cameras can be used to learn where the occupants are and what they do, they raise privacy concerns. While this can be mitigated by severely reducing camera resolution, it will also impact the utility of the camera network. This work investigates and quantifies the tradeoff between camera resolution and action recognition accuracy. Rather than building a physical testbed to carry out this study, we use a graphics engine to simulate a room with 5 cameras, and to animate avatars using skeletal movements of real users captured by a Kinect v2 camera. We study resolutions from 100×100 pixels down to 1×1 using a state-of-the-art action recognition method at higher resolutions and we propose a new approach at ultra-low resolutions. In extensive simulations, we conclude that on a dataset of 12 individuals performing 4 actions our algorithm applied to single-pixel data performs very close to the state-of-the-art method applied to 100×100 data, suggesting that reliable action recognition can be achieved without compromising occupant's identity.

Index Terms— Action recognition, privacy-preserving, very low resolution

1. INTRODUCTION

While many of our personal items are already "smart" (e.g., smartphones, smart watches), extending this to the infrastructure around us is more challenging. One effort in this direction is research on smart spaces – environments that allow intelligent interaction with their occupants, be it a living or conference room. Among the promised benefits of future smart rooms are improved energy efficiency, sustainable health benefits and increased productivity. For instance, localization of human subjects may enable direct illumination of target areas saving energy in areas void of humans. Recognizing the types of activities may allow task-optimized lighting, e.g., reduced screen glare when working on a laptop. As for productivity, localization of occupants may help maximize throughput rates in visible light communication (VLC) between fixed transceivers (ceiling, walls) and mobile devices (smartphones, tablets, laptops). Finally, hand gestures can be used to control various room conditions (e.g., temperature, light). Realizing these benefits, requires, among other things, reliable recognition of human actions and activities.

Fig. 1. A smart room simulated in Unity3D© with 5 cameras and a single omni-directional light source mounted overhead.

Although extensive research has been performed to date on human action recognition, most of the work exploits video cameras. However, with the increasing concern about privacy, standard video cameras seem unsuitable for smart spaces of the future. Concerns about privacy can be partially addressed by significantly reducing the camera resolution. This, however, degrades recognition accuracy. While this can be mitigated by using multiple sensors it is unclear to what extent, thus motivating a study of tradeoffs between camera resolution and action recognition accuracy. Our ultimate goal is to evaluate whether extreme "single-pixel" cameras can provide enough information to accurately recognize actions.

Our study would be impractical on a real testbed; changing the number of cameras, and their resolutions, locations, orientations for each studied scenario is very tedious. Therefore, we chose to animate avatars in a simulated smart room using the Unity3D© graphics engine. However, we animated the avatars with *true* human motions captured by Kinect v2 camera in a physical testbed. Fig. 1 shows our simulated smart room from one viewpoint.

In this paper, we report results of our simulation using 5 ceiling-mounted cameras (Fig. 1) with resolutions varying

*This material is based on work supported by the NSF under Smart Lighting ERC Cooperative Agreement No. EEC-0812056.

from 100×100 at the high end to 1×1 at the extreme low end (5 single-pixel cameras). Since at the low end, standard action recognition methods based on optical flow, trajectories or silhouettes cannot be used, we propose a new approach based on the time series of brightness values.

2. RELATED WORK

There is an excellent review of human activity recognition using full resolution color cameras by Aggarwal and Ryoo [1]. Arguably, the most important step in such methods is feature extraction from the data produced by video cameras. Some of the most common features used are optical flow [2], point trajectories [3], silhouettes [4, 5, 6, 7] and spatio-temporal interest points [8, 9]. Much less research has been devoted to action recognition from low-resolution data. Efros *et al.* [2] apply optical flow when the human is at least 30 pixels tall, and use nearest-neighbor classification of optical flow sequences with a distance metric based on frame-to-frame normalized correlation aggregated in time. Ahad *et al.* [10] use time-aggregated optical flow to obtain directional motion history images. They extract Hu moments from each channel to obtain the feature vectors and apply a K-nearest neighbor classifier. This method works well on resolutions down to 32×24. Reliable optical flow cannot be computed at lower resolutions. Tao *et al.* [11] use a network of 20 ceiling-mounted binary infrared sensors to recognize daily activities without privacy violation. An SVM classifier applied to the short-duration averages of binary values gives a good recognition performance, however we note that there is a significant location bias in most of the activities thus potentially inflating the recognition performance. The aforementioned features tend to perform poorly at low resolutions. At resolutions from 10×10 down to 1×1, these features are either ill-posed or too noisy. In addition, privacy-preserving tracking and coarse pose estimation (estimating standing or sitting) have been proposed by Jia and Radke [12] based on a sparse set of time-of-flight measurements.

3. OVERVIEW OF OUR APPROACH

Characterizing the full range of possible trade-offs between camera resolution and recognition accuracy is a daunting task since there are simply too many degrees of freedom available to explore: the number of cameras, their zoom, and color settings, their position and orientation relative to the room, and the number, position, and orientation of training and test avatars in the room. In this work, we explore only a small sliver of this problem by focusing on uncovering the trade-off between camera resolution and action recognition accuracy while holding all other degrees of freedom fixed. In particular, we study a network of 5 grayscale cameras at fixed positions and orientations relative to the room. We further assume that there is a single avatar in the room and the zoom setting of each camera is such that the avatar roughly covers its field of view (FOV). 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 resolution. With this setup, we change the resolution setting of the cameras from 100×100 all the way down to 1×1 and evaluate the action recognition accuracy.

Action Recognition at Higher Resolutions: In order to study the trade-off between resolution and accuracy, we need an algorithm for action recognition. Instead of developing a completely new algorithm, we use the method proposed in [7] whose performance is competitive with current state-of-the-art methods. This method captures the shape of the silhouette tunnel of an action - which can be obtained via background subtraction - through the empirical covariance matrix of certain local features defined at every pixel inside the tunnel. These local features include normalized spatio-temporal coordinates of the pixel and distances from the pixel to the tunnel boundaries along a fixed set of directions (see [7, Sec. IV.A] for a detailed description). The similarity between two action sequences is then measured by the ℓ_1 distance between their scale-normalized feature covariance matrices. The action recognition algorithm is then a simple nearest-neighbor classifier based on this distance metric.

Dealing with Extremely Low Resolutions: The success of current state-of-the-art action recognition algorithms hinges on reliable estimation of certain basic features such as silhouette tunnels or optical flow. At resolutions roughly below 10×10, such features cannot be reliably or even meaningfully estimated. For example, the silhouette tunnel will have imprecise pixel-to-boundary distance features thus causing a drastic drop in performance. While more nuanced approaches are certainly possible to deal with such low resolutions, in this work we use a simple nearest-neighbor classifier based on the ℓ_1 distance between the mean-subtracted interpolated time-series of pixel grayscale values. To elaborate, let $x_{i,j,k}[t]$ and $y_{i,j,k}[t]$ denote the grayscale values of pixel (i, j) in camera k at time t for action sequences x and y, respectively, after they have been linearly stretched to a common length of $T = 500$ using cubic-spline interpolation. Let

$$\mu^x_{i,j,k} = \frac{1}{T}\sum_{t=1}^{T} x_{i,j,k}[t], \;\; \mu^y_{i,j,k} = \frac{1}{T}\sum_{t=1}^{T} y_{i,j,k}[t]$$

denote the time-averaged mean grayscale values of x and y respectively at pixel (i, j) in camera k (shown in Fig.1). Then, we measure the distance between x and y as

$$d(x,y) = \sum_{k=1}^{5}\sum_{t=1}^{T}\sum_{i=1}^{R}\sum_{j=1}^{R}\left|x_{i,j,k}[t] - \mu^x_{i,j,k} - y_{i,j,k}[t] + \mu^y_{i,j,k}\right|$$

where the camera resolution is $R \times R$. We subtract the mean because it represents a "static" pixel- and camera-dependent illumination which does not capture action dynamics. The

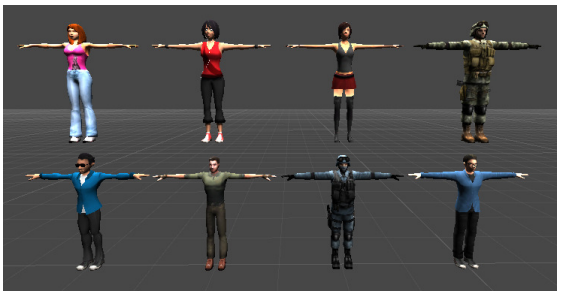

Fig. 2. Avatars used in experiments

advantage of this approach is its simplicity. Admittedly more sophisticated fusion methods are certainly possible, but in our approach, distances from all 5 cameras are combined additively with equal weights. However, we should expect this metric to be sensitive to positions of avatars in the FOV and indeed this is reflected in our experimental results.

4. DATASET

In order to empirically quantify the trade-off between camera resolution and action recognition accuracy, we created a dataset consisting of 4 actions, each repeated 3 times, from 12 different subjects (5 female, 7 male), comprising young to middle-aged adults, using a single Kinect v2 camera facing the subjects. We selected 4 actions: *Answering Phone* (about 12 sec long), *Showing of Hand* (akin to voting – 6 Sec), *Writing on Board* (10 sec), and *Walking* (4 sec). These are representative of actions that can occur in a typical seminar-room environment. All actions are performed while standing. Walking is the only one where there is significant translational motion of the subject within the FOV.

To simulate the actions in a virtual room environment with arbitrary camera viewpoints and resolutions, we imported the skeletal data sequences containing the dynamics of actions into Unity3D©. We used 8 avatars available in Unity3D©, 5 male and 3 female (Fig. 2), and animated them using the imported skeletal coordinates. With the exception of *Walking*, the avatars are situated in the same location to remove action-specific location-bias which can artifically boost action recognition accuracy. We placed 5 cameras on the ceiling of the simulated room (Fig. 1). Views of actions from the same camera are shown in Fig. 3 while those from all 5 cameras for *Showing of Hand* are shown in Fig. 4.

5. EXPERIMENTAL RESULTS

In order to quantify the tradeoff between resolution and action recognition accuracy, for each resolution setting we compute the average Correct Classification Rate (CCR) using a variation of leave-one-out cross-validation (LOOCV) as follows. We compute the average CCR score across M iterations where each iteration corresponds to a random assignment of same-gender avatars to each of the 12 subjects in our

Answering Phone *Showing of Hand* *Writing on Board* *Walking*

Fig. 3. Sample frames from 4 actions for avatar 3.

Camera 1 Camera 2 Camera 3 Camera 4 Camera 5

Fig. 4. 5 viewpoints for *Showing of Hand* by avatar 6.

Table 1. CCR for time-series and silhouette-covariance methods across a range of resolutions using 5 cameras jointly.

Time-series method (M=100)			Silh.-cov. method (M=1,000)		
$R \times R$	CCR	StDev	$R \times R$	CCR	StDev
10×10	45.26%	3.33%	10×10	84.47%	1.78%
5×5	69.44%	3.31%	20×20	84.90%	1.78%
4×4	80.14%	3.24%	30×30	84.90%	1.92%
3×3	85.56%	3.68%	40×40	85.87%	1.68%
2×2	87.04%	2.76%	50×50	85.70%	1.67%
1×1	**85.69**%	3.16%	100×100	**86.88**%	1.70%

dataset. Specifically, in each iteration, each of the 12 subjects is sequentially selected to be the test subject. Then, one same-gender avatar is randomly assigned to the test subject and the assigned avatar and all 12 samples from that subject (4 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). Then, each of the 12 samples of the test avatar is classified using the 132 samples of the 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 iterations to obtain the final average CCR. We perform this test on resolutions from 10×10 down to 1×1 and report the results in Table 1.

The single-pixel CCR is 85.69% and increases to 87.04% for 2×2 resolution. This is to be expected as more data are available at the higher resolution. However, the CCR drops at subsequent resolutions with a severe dip at 5×5 to 69.44%. This was unexpected but upon careful analysis we realized that at higher resolutions our simplistic metric between two time-series does not account for differences in the locations, sizes, and orientations of the avatar projections under comparison. At 1×1 resolution, this is not an issue of course.

In order to verify this hypothesis, we performed a coarse alignment by co-locating the centroids of silhouettes (easily extracted by background subtraction). In consequence, at 10×10 resolution CCR increased from 45.26% to 82.95% and at 5×5 resolution it went from 69.44% to 76.17%. Clearly, the time-series method is alignment-sensitive but, fortunately, not so much at extremely low resolutions.

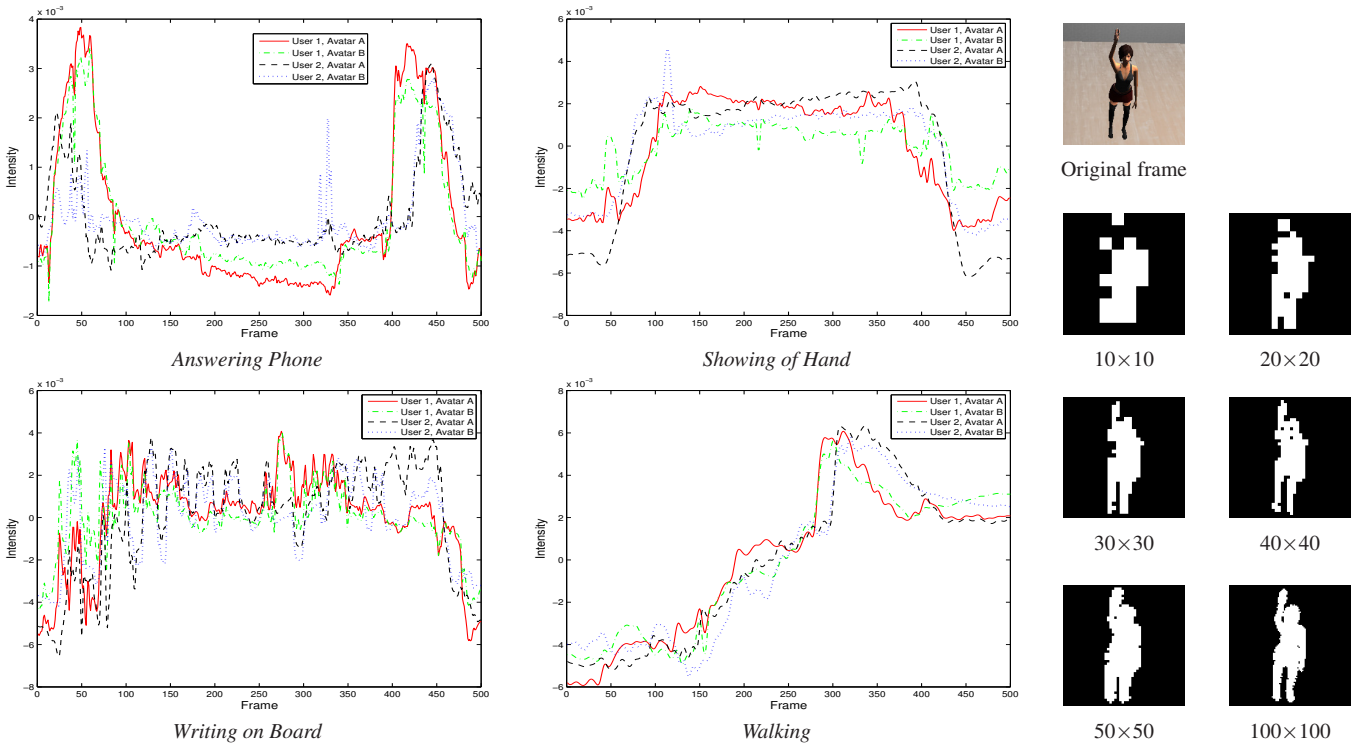

Fig. 5. Left: Single-pixel signals for 4 actions, 2 users and 2 avatars. Right: Sample silhouettes extracted at different resolutions and then re-scaled to the same size.

In Fig. 5, we show examples of signals obtained from a single-pixel camera for all 4 actions performed by 2 different users and each rendered by 2 different avatars. Clearly, *Showing of Hand* and *Writing* exhibit similar profiles and are indeed the main source of confusion for our method (we omit the confusion matrix due to space constraints).

Table 2. CCR for the proposed time-series method (M=100) using each of the 5 cameras separately (Fig. 4).

	Cam1	Cam2	Cam3	Cam4	Cam5
1×1	70.44%	78.26%	74.81%	58.72%	55.27%
2×2	76.81%	78.67%	80.78%	69.49%	65.65%
3×3	77.83%	72.60%	73.70%	73.57%	71.92%
4×4	72.04%	55.25%	65.72%	69.72%	76.20%
5×5	61.32%	64.38%	67.99%	70.01%	77.35%
10×10	60.00%	57.82%	59.31%	59.44%	68.87%

Table 3. CCR for the silhoutte-covariance method (M=1,000) using each of the 5 camera separately (Fig. 4).

	Cam1	Cam2	Cam3	Cam4	Cam5
10×10	81.50%	83.92%	81.85%	81.62%	80.26%
20×20	81.96%	82.47%	82.19%	81.77%	84.28%
30×30	81.48%	82.57%	82.40%	84.33%	86.05%
40×40	81.97%	84.60%	83.36%	84.84%	86.65%
50×50	81.97%	85.50%	83.83%	85.20%	86.88%
100×100	83.21%	87.42%	86.03%	83.97%	85.37%

Results for the silhouette-covariance method at resolutions 10×10 and higher are also shown in Table 1. Although the CCR drops consistently with resolution reduction, which was expected due to the degradation of silhouettes (Fig. 5), interestingly this drop is rather small: from 86.88% at 100×100 to 84.47% at 10×10. Furthermore, despite a relatively accurate silhouette at 100×100 the covariance method's CCR is only 1.19% higher than that of the time-series method on single-pixel data. In Tables 2 and 3, we compare camera-by-camera performance of the two methods (the computation of distances is confined to the output of one camera). While the CCR for the silhouette-covariance method is quite consistent across cameras and resolutions, the performance of the time-series method exhibits wide variations across both. Aggregating time-series distances across cameras largely compensates for the variability of single-camera outcomes.

6. CONCLUDING REMARKS

This work studied and quantified the impact of camera resolution on action recognition accuracy in a simulated environment environment (Unity3D© + Kinect v2). Results for a dataset of 12 individuals performing 4 actions indicate, somewhat surprisingly, that the recognition accuracy at single-pixel resolution can be quite close to that at 100 × 100 resolution. This work has explored just one degree of freedom in the space of possible trade-offs between recognition accuracy and resolution. The proposed approach can be used to study and quantify trade-offs for other parameters such as the number of cameras, their zoom settings, their position and orientation relative to the room, etc.

More information on this research and dataset, as well as some resources are available on-line [13].

7. REFERENCES

[1] J.K. Aggarwal and M.S. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 16:1–16:43, 2011.

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

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

[4] M.-Ch. Roh, W. J. Christmas, J. Kittler, and S.-W. Lee, "Robust player gesture spotting and recognition in low-resolution sports video," in *European Conference on Computer Vision (ECCV)*, 2006.

[5] L. Wang and D. Suter, "Learning and matching of dynamic shape manifolds for human action recognition," *IEEE Transactions on Image Processing*, vol. 16, no. 6, pp. 1646–1661, 2007.

[6] B. Saghafi and D. Rajan, "Human action recognition using pose-based discriminant embedding," *Signal Processing: Image Communication*, vol. 27, no. 1, pp. 96–111, 2012.

[7] K. Guo, P. Ishwar, and J. Konrad, "Action recognition from video using feature covariance matrices," *IEEE Transactions on Image Processing*, vol. 22, no. 6, pp. 2479–2494, 2013.

[8] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *International Workshop on Performance Evaluation of Tracking and Surveillance, ICCV*, 2005.

[9] 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.

[10] M.A.R. 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*, vol. 3, no. 1, pp. 17–24, 2010.

[11] S. Tao, M. Kudo, and H. Nonaka, "Privacy-preserved behavior analysis and fall detection by an infrared ceiling sensor network," *Sensors*, vol. 12, no. 12, pp. 16920–16936, 2012.

[12] L. Jia and R.J. Radke, "Using time-of-flight measurements for privacy-preserving tracking in a smart room," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 1, pp. 689–696, Feb. 2014.

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