Supplementary material for ‘Reconstruction-free action inference from compressive imagers’

Kuldeep Kulkarni, Pavan Turaga

1 Reconstruction-free recognition on KTH dataset

The KTH dataset contains 2391 sequences of six different actions: ‘boxing’, ‘hand-clapping’, ‘hand-waving’, ‘jogging’, ‘walking’ and ‘running’, performed by 25 actors under 4 different scenarios, s1, s2, s3 and s4. We conducted two separate recognition experiments using 5-fold cross validation for a fixed compression ratio of 100. In the first experiment, we used ‘Type-1’ filters and in the second, we used ‘Type-2’ filters. ‘Type-2’ filters are required to be transformed to the test video’s viewpoint before calculating responses. As expected, the recognition accuracy (shown table 1) was higher when ‘Type-2’ filters were used, thus corroborating the need to transform all the training sequences to the same viewpoint before the filter is synthesized. Overall recognition accuracy for the second experiment is 74.76%, which is comparable to that obtained using uncompressed MACH filtering as in [1]. The confusion table for the second experiment for scenario 1 is given in table 1.

Activity	Run	Jog	Walk	Boxing	Wave	Clap
Run	**0.84**	0.12	0.01	0	0	0
Jog	0.08	**0.85**	0.07	0	0	0
Walk	0.02	0.15	**0.83**	1	0	0
Boxing	0	0	0	**0.84**	0.01	0.14
Wave	0	0	0	0.03	**0.94**	0.03
Clap	0	0	0	0.08	0.1	**0.81**

TABLE 1
Confusion table for experiment II at a compression factor = 100 for scenario 1 of KTH database. Overall recognition rate for this scenario is 85.17 %.

Scenarios	s1	s2	s3	s4	Overall
Experiment I	78.401	55.10	63.50	71.33	67.08
Experiment II	85.17	60.10	73.91	79.97	74.76
Oracle MACH [1]	NA	NA	NA	NA	80.90

TABLE 2
KTH dataset: The recognition rate for experiment I and II at compression rate of 100. We note that using ‘Type 2’ filters for translational actions is better, thus corroborating the need for transforming the training set to a single viewpoint. The overall recognition accuracy for experiment II is comparable to that of uncompressed MACH filtering in [1].

2 Weizmann dataset: Spatial localization of action from compressive cameras without reconstruction

Figure 1 shows action localization in a few frames for the ‘One handed wave’ and ‘Jack’ action.

3 UCF sports dataset

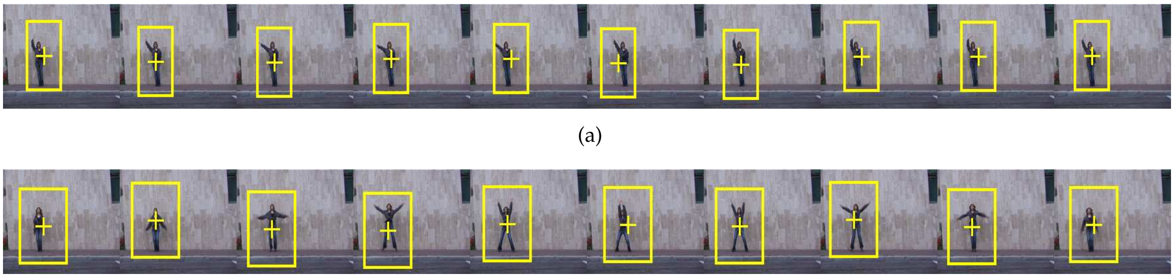
Confusion table at compression ratio = 300: The confusion matrix for compression ratio 300 is shown in table 3.

Spatial Localization of Actions from Compressive Cameras without Reconstruction: Figure 2 shows action localization for some correctly classified instances across various actions in the dataset, for Oracle MACH and compression ratio = 100. It can be seen that action localization is estimated reasonably well despite large scale variations and extremely high compression ratio.

References

[1] M. D. Rodriguez, J. Ahmed, and M. Shah, “Action MACH: a spatio-temporal maximum average correlation height filter for action recognition,” in *IEEE Conf. Comp. Vision and Pattern Recog*, 2008.

- *K. Kulkarni and P. Turaga are with the School of Arts, Media and Engineering and School of Electrical, Computer and Energy Engineering, Arizona State University. Email: kkulkar1@asu.edu, pturaga@asu.edu.*



(a)

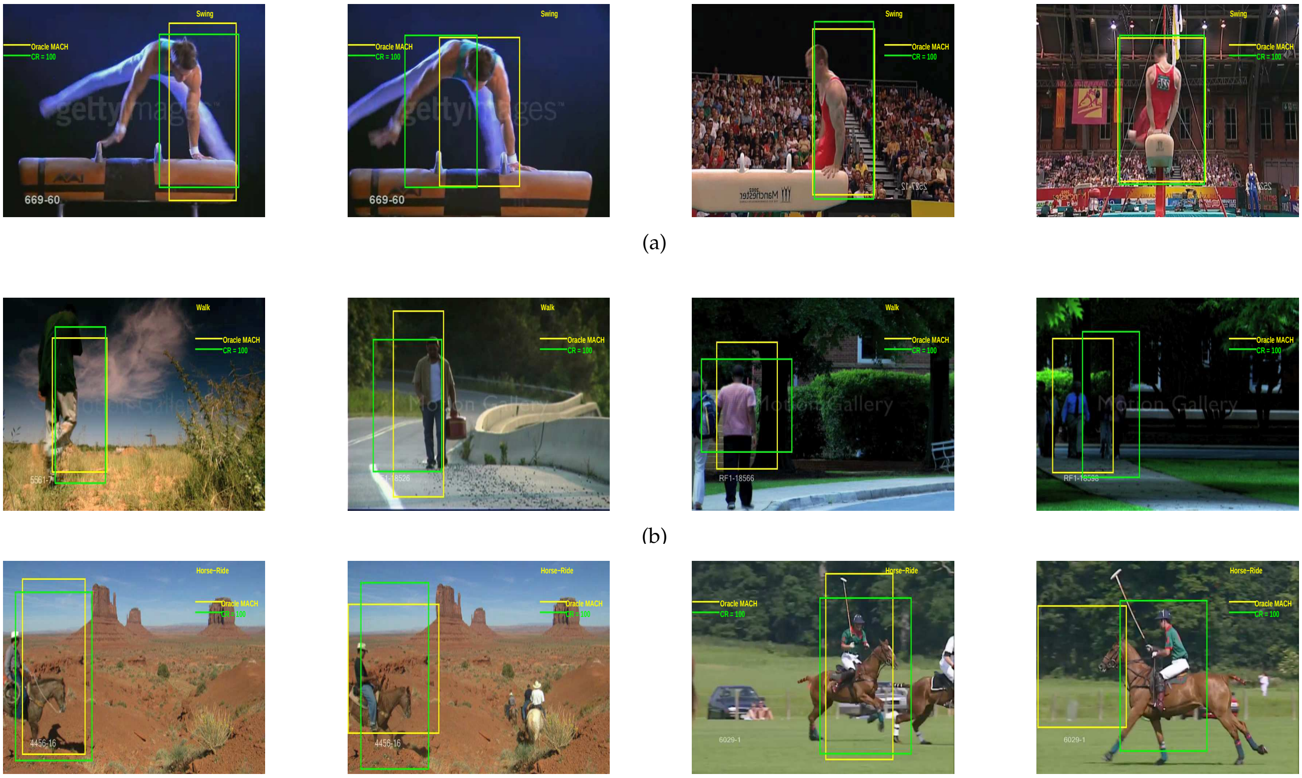
(b)

Fig. 1. Spatial localization of subject without reconstruction at compression ratio = 100 for different actions in Weizmann dataset. a) One handed wave. b) Jack

Action	Golf-Swing	Kicking	Riding Horse	Run-Side	Skate-Boarding	Swing	Walk	Diving	Lifting
Golf-Swing	**55.56**	0	27.78	0	0	5.56	11.11	0	0
Kicking	0	**95**	0	5	0	0	0	0	0
Riding Horse	0	0	**75**	16.67	0	8.33	0	0	0
Run-Side	0	0	7.69	**38.46**	7.69	30.77	7.69	7.69	0
Skate-Boarding	8.33	0	0	8.33	**50**	16.67	16.67	0	0
Swing	0	0	0	12.12	12.12	**72.73**	3.03	0	0
Walk	0	0	0	0	4.55	22.73	**72.73**	0	0
Diving	0	0	14.29	14.29	0	14.29	0	**57.14**	0
Lifting	0	0	0	0	0	0	0	16.67	**83.33**

TABLE 3

Confusion matrix for UCF sports database at a compression factor = 300. Recognition rate for this scenario is 68 %.



(a)

(b)

(c)

Fig. 2. Reconstruction-free spatial localization of subject for Oracle MACH (shown as yellow box) and STSF (shown as green box) at compression ratio = 100 for some correctly classfied instances of various actions in the UCF sports dataset. a) Swing b) Walk c) Horse. Action localization is estimated reasonably well directly from CS measurements even though the measurements themselves do not bear any explicit information regarding pixel locations.