Correcting Cuboid Corruption For Action Recognition In Complex Environment

Syed Zain Masood Adarsh Nagaraja Nazar Khan Jiejie Zhu Marshall F. Tappen
University of Central Florida
Orlando, Florida

Abstract

The success of recognizing periodic actions in single-person-simple-background datasets, such as Weizmann and KTH, has created a need for more difficult datasets to push the performance of action recognition systems. We identify the significant weakness in systems based on popular descriptors by creating a synthetic dataset using Weizmann dataset. Experiments show that introducing complex backgrounds, stationary or dynamic, into the video causes a significant degradation in recognition performance. Moreover, this degradation cannot be fixed by fine-tuning the system or selecting better interest points. Instead, we show that the problem lies at the cuboid level and must be addressed by modifying cuboids.

1. Introduction

For action recognition in complex environments, knowing where in the video the action is being performed is very useful as it helps prune out irrelevant background movement. Such accurate localization is not easily obtained, especially in the presence of substantial background motion. Assuming we have near-perfect localization, one expects that this localization information would make it relatively easy to improve the recognition performance of an action recognition system. Moreover, one might also believe that this localization information should make it straightforward to achieve results nearly as good as would be achieved on videos with stationary backgrounds.

In this paper, we show that *that is not necessarily the case*. We show that using localization to improve action recognition with a bag-of-words recognition system is a surprisingly subtle problem. We focus on bag-of-words systems, where the classifier is based on quantizing image descriptors gathered at interest points and examining the frequency of different types of descriptors, because they have been a very popular strategy in action recognition [3, 18, 16, 17, 11].

Our message for bag-of-words systems can be summarized in the following four statements:

1. Simply using the localization information to prune irrelevant interest points will not achieve the best results possible.

2. Even perfect localization for interest point pruning cannot achieve the best results possible.

3. Complex background motion reduces classification accuracy because the cuboids themselves are corrupted by the background motion. Systems that do not address this corruption will be limited in performance.

4. Fortunately, even with inaccurate, automatic localization, the effects of cuboid corruption can be ameliorated with simple modifications. Combined with interest point pruning strategies, a system can perform equally well on simple as well as complex datasets.

Even though localization strategies for interest point pruning solve the action recognition problem on simple datasets, the same cannot be said about complex datasets. Pruning is helpful in eliminating erroneous background interest points, but it fails to deal efficiently with irrelevant background information within selected interest point cuboids. Current action recognition systems [12, 15, 2, 9] fail to address this concern.

1.1. Paper Organization

Following is a brief overview of how the paper is organized:

- To understand how background complexity affects recognition accuracy, we create a new synthesized dataset that contains videos of simple actions on complex background (refer to Section 2). Using this dataset makes it easier to analyze how simply modifying background complexity influences results.

- We present our basic classifier method and show that it performs as well as state-of-the-art on well known datasets (refer to Section 3). However, in Section 4, we discuss why it fails to perform equally well on the new synthesized dataset.

2011 IEEE International Conference on Computer Vision Workshops
978-1-4673-0063-6/11/$26.00

- In Section 5, we show how localization is imperative for achieving improved results on this dataset. Even using average automatic localization, we show how simple *but effective* techniques like interest point pruning and correcting cuboid corruption lead to a significant improvement in results.
- Section 6 shows that our system works on commonly used UCF Sports action datasets.

2. Constructing a New Dataset to Understand the Effect of Background Complexity

In order to understand the effect of background complexity on recognition performance, we create a synthetic dataset with the aim of *isolating* the effects due to complex backgrounds. We do so by constructing synthetic videos of the same action being performed on different complex backgrounds. This way the difference in videos comes only from the background complexity.

To maintain focus on the problem of recognizing specific actions, we introduce a new synthetic complex dataset based on the Weizmann [1] dataset. This dataset is constructed by extracting action masks, provided on-line [1], for each Weizmann dataset video and then replacing the background with a randomly selected Youtube video.

In establishing our reasoning for the construction of a new dataset, it is helpful to first consider the key properties of the Weizmann dataset. It contains a single actor performing simple periodic actions with simple fixed backgrounds. This construction forces the recognition system to focus directly on recognizing the action being performed by the actor. Also, the dataset allows us to control the quality of localization of the action being performed.

For this new synthesized dataset, the central recognition problem remains the same, but the task is made more difficult by the addition of the complex background. Essentially, our goal is to only modify one aspect, the background, during the recognition experiments.

2.1. Construction Choices

The Weizmann dataset was chosen because the actions are simple and coherent. In addition, each video has an associated action mask which makes it possible to extract the action and construct new videos with complex backgrounds.

We avoid the use of realistic complex datasets like Youtube [17, 15] and Hollywood [12, 18] because isolating the effect of background complexity from within the highly complex structure (multiple people, multiple actions, camera movement, high diversity within action class) of these datasets is extremely challenging.

We chose not to make a similar construction for the KTH dataset because the running and jogging actions in that dataset have not been recorded perfectly. Recent action recognition systems have near 100% accuracy on all actions except jogging and running [7, 17, 15, 12]. This is because the difference between these actions is not discernible for portions of this dataset, such as the videos from person 2.

To justify this decision, we conducted an experiment, involving humans, to gauge the difficulty of correctly recognizing actions between jogging and running. Each person was shown 2 training videos of each jogging and running and then was asked to correctly label a total of 50 test videos. Human subjects were only able to correctly recognize 90% of the jogging and running videos shown. This is approximately the same accuracy as the state-of-the-art. The difficulty that humans have with running and jogging in this set makes it less desirable for evaluating machine vision systems.

2.2. Construction Methods

We create a new dataset using Weizmann dataset action masks and background from Youtube videos. We downloaded a total of 15 Youtube videos making sure that each of them contain some complex scene. We then randomly select a Youtube video from this pool and perform matting with one of the Weizmann dataset action mask. Keeping the Youtube video pool considerably lower than the number of action masks (93 in this case) ensures different actions being performed on the same background and thus diminishing the role of background in differentiating actions.

The dataset is developed using the following strategy:

- **UCF Weizmann Dynamic** The whole video is matted with the action mask (Refer to Figure 2). The moving background makes it a much harder problem to recognize actions. This helps to analyze how increased background complexity affects recognition.

This new dataset will be made public and provided online [2]. It should be noted that when creating the UCF Weizmann Dynamic dataset, we make sure that none of the Youtube backgrounds have humans in it. This is a necessity as the presence of humans in background videos is most likely to be accompanied by some action, leading to multiple actions in a single video. Our aim is to isolate the effect of background motion as opposed to multiple human actions and therefore we avoid using background videos with humans in them.

Our methodology of creating a complex dataset for simple actions is different from [14]. Our synthesized dataset is complete replica of the simple dataset in terms of the action being performed, and accuracy of the recognition can be compared directly. Since, we use matting [13] to create new dataset, it will not add any biases, due to change

[1] http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html

[2] http://www.cs.ucf.edu/~smasood/datasets/UCFWeizmannDynamic.zip

Figure 1. Examples of the Weizmann (top row) and UCF Weizmann Dynamic (bottom row) datasets. Each video in the UCF Weizmann Dynamic dataset has the moving complex background. This indicates the background complexity of gradients, textures and contrasts on which the actions are overlayed.

Figure 2. Examples of the UCF Weizmann Dynamic dataset. The figure shows frames 1, 11, 21, 31 and 41 of 2 running actions with complex, dynamic backgrounds. The top row indicates running action overlayed on a backround video with fast moving trees with high gradients and textures. Bottom row indicates running action overlayed on a slow moving eagle video. Care was taken not to have background videos with humans in order to isolate the effect of background motion as opposed to multiple human actions.

in the actor performing the action. Because of the synthetic construction of this dataset, matting artifacts could pose an issue and this is discussed next.

2.2.1 Addressing Matting Artifacts

To measure the effect of matting artifacts, we constructed a separate dataset by matting the Weizmann dataset action masks with a simple static gray background. We found negligible ($\approx 2\%$) change in performance, making us confident that matting artifacts were not an issue.

3. Baseline Method and Performance

Having created this new synthesized dataset, we need to decide on a baseline system to be used. In this section, we explain the basic classifier approach we adopted and later evaluate its performance on both simple and complex datasets.

3.1. Baseline: Basic Bag-of-Features Classifier

We use a standard bag-of-features approach [3] as our baseline method. We make use of the code provided online[1]. Given any video sequence, we detect spatio-temporal interest points, extract cuboids centered around the interest points and compute gradient descriptors histograms. These histogram of gradients (HoG) are concatenated and Principal Component Analysis (PCA) is applied to project the gradients into lower dimensional space. Visual vocabulary is constructed using subset of the dataset followed by histogram representation formation for each video sequence. For classification, Support Vector Machine (SVM) classifier [2] is learnt using histogram intersection kernel and testing is

[1] http://vision.ucsd.edu/~pdollar/toolbox/doc/
[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm

Dataset	Our Baseline (Section 3)	STIPS (HOF) [12]	Liu et al. [15]
Weizmann	98%	92%	91%
KTH	93.5%	92%	93.8%
Youtube	65%	–	71.2%

Table 1. Comparison of our baseline and other state-of-the-art techniques on well known datasets. Comparable results on KTH and Youtube datasets shows robustness of our baseline approach.

Dataset	Our Baseline (Section 3)	STIPS (HOF) [12]
Weizmann	98%	92%
UCF Weizmann Dynamic	36.5%	31%

Table 2. Comparison of our baseline and other state-of-the-art technique on the synthesized dataset. We observe a significant drop in performance when switching from Weizmann dataset to the new UCF Weizmann Dynamic dataset.

done using leave-one-out-cross-validation (LOOCV).

Since the Weizmann dataset is relatively small, most research studies use the video reflection technique to double the size of the dataset [19]. This involves horizontally flipping each video and saving it as a new video. We use the same reflection approach for all our datasets.

The performance of this basic bag-of-features classifier as well as that of the state-of-the-art [12, 15] on different datasets is shown in Table 1. Despite being a simple technique, our baseline method performs reasonably well and is robust across different known datasets.

In the next section, we will discuss why performance degrades for these new synthesized complex datasets and what measures can be taken to improve results. Derived solutions are later tested on a realistic action dataset i.e. UCF Sports.

4. Measuring Performance Degradation

Having evaluated our basic classifier system on well known datasets, we now focus on how the system performs on the new synthesized dataset. Table 2 shows a comparison of the Weizmann and UCF Weizmann Dynamic datasets for our baseline system. We observe a sharp drop in accuracy when switching from Weizmann dataset to the newly synthesized UCF Weizmann Dynamic dataset. Since the actions are exactly the same for both datasets, it is only logical to assume that the performance degradation is caused by the increased background complexity in the new UCF Weizmann Dynamic dataset.

Before devising a new solution, we first try some of the well known strategies in order to achieve improved results. The next section details these methods and shows how they fail to solve the posed problem.

4.1. Unsuccessful Strategies For Dealing With Performance Degradation

A general approach towards solving this degradation in performance is to fine tune the system parameters. For this reason, we experimented using:

- different vocabulary sizes of 250, 500 and 1000 clusters
- averaging of features across different temporal scales [12, 6]
- cleaner vocabulary generated for Weizmann dataset
- χ^2 kernel for SVM classification [20]

We achieved a maximum improvement of 2% using these techniques, thus failing to solve the particular problem that we pose here – recognition with complex backgrounds.

Background complexity plays a vital role when recognizing actions in videos. Even if the actions are simplistic, recognition systems performance is heavily dependent on the background they are performed on. In the next section we will discuss how the use of action localization goes a long way in rectifying this problem. It is no surprise that localization is helpful but, as will be shown below, it is the application of localization that is equally important.

5. Utilizing Action Localization For Handling Performance Degradation

We observed that the introduction of complex background in videos of simple actions greatly affects recognition performance (refer to Table2). Since the only change between the Weizmann and UCF Weizmann Dynamic datasets is of the background, it is reasonable to say that the drop in accuracy is only due to the change in background complexity. This is because increased background complexity leads to detection of irrelevant background interest points that are a main source of performance degradation. One would assume that eliminating these background interest points should solve the problem. However, *that is not the case*. In fact, it is the use of localization for both pruning irrelevant interest points and eradicating background corruption inside cuboids that leads to optimal results. Thus we can say that:

- Action localization is important but
- Application/use of localization is equally significant

We propose a stepwise solution to the above posed problem:

- First and foremost, we need a good automatic action localization methodology (preferably a tight bounding box around the person performing the action).

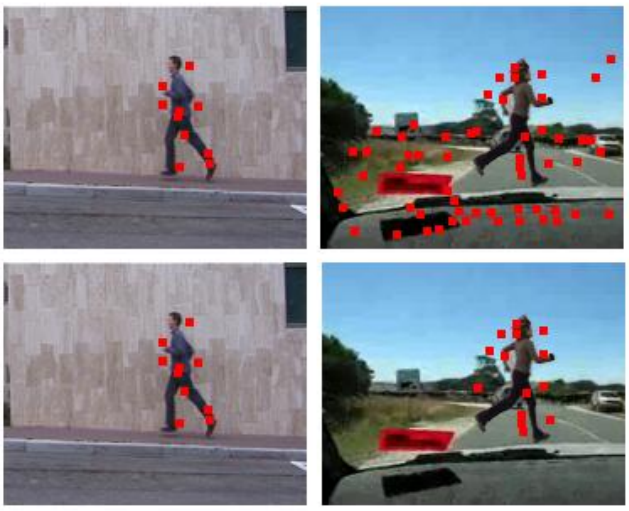

Figure 3. Top row shows the interest points without pruning for Weizmann and UCF Weizmann Dynamic datasets respectively. Bottom row shows the interest points for the same frame after pruning. For better recognition, it is thus important to remove background interest points.

- Once we have localization information, we eliminate all interest points detected due to background motion
- Having removed erroneous interest points, we use localization to correct cuboid corruption due to background information i.e. mask out background pixel values within valid cuboids.

Below, we will discuss each of the above strategies in detail. We will show how simply localizing the action and pruning irrelevant interest points is insufficient and that optimal results are achieved only when localization is directly used to modify the cuboids. Thus, these experiments will show that systems like [2, 9, 10] that use localization just to eliminate irrelevant interest points will have inferior performance compared with a system that uses localization information to also directly modify the cuboids.

We will build on the baseline system described in Section 3. To gauge performance of our system and to provide an upper bound on achievable accuracy, we will also present results obtained using ground-truth localization. Ground-truth localization masks are generated by forming a tight bounding box around the silhouette mask, available with the Weizmann dataset, at each frame.

Having analyzed and proposed solutions to the posed problem, in Section 6 we will show results on the UCF Sports dataset which is a commonly used complex action dataset in the vision community.

5.1. Automatic Localization

Since adding background complexity leads to significant increase in false positive interest point detections, it is imperative to design a system that accurately detects regions where the action is being performed. This is especially important for the UCF Weizmann Dynamic dataset where there is significant background motion. Once we have good localization of the action, discarding irrelevant interest points and modifying cuboids can be easily implemented. In reality however, such localization is hard to achieve for realistic datasets.

Method	UCF Weizmann Dynamic
Our Baseline (Section 3)	36.5%
Automatic Localization + Interest Point Pruning	41%
Ground-truth Localization + Interest Point Pruning	68%

Table 3. The above table shows the accuracy on synthesized complex dataset when using interest point pruning with automatic localization. Best possible results for interest point pruning with ground-truth localization are also shown. Although results improve, they are still not comparable to those achieved on Weizmann dataset using our baseline system (Table 2).

We combine an off-the-shelf human detection system [5, 4] and a saliency detection method [8] for obtaining automatic localization information of the action being performed. We employ the same technique when dealing with realistic UCF Sports dataset in Section 6.

5.2. Interest Points Pruning

Directly running our baseline system on the new synthesized dataset results in interest points detected due to both the action and background motion. Having computed automatic localization information, we can now remove irrelevant background interest points. The goal is to discard all interest points lying outside the automatic localization mask calculated previously. This technique is applied at each frame of the action video sequence. With the removal of these background interest points, the recognition performance is expected to improve.

Figure 3 shows the interest points generated for the mentioned dataset. We see that almost all interest points in the Weizmann dataset are on or near the person performing the action. For the UCF Weizmann Dynamic dataset however, a significant number of interest points are due to background motion. It is essential that we remove these interest points for improved recognition accuracies. We thus prune interest points lying outside the automatic localization masks generated for this dataset. It should be noted that these localization masks are in fact rectangular bounding boxes and so different from silhouette masks. After pruning, interest points for the Weizmann dataset remain the same. However, interest points from the UCF Weizmann Dynamic dataset are reduced by large extent (see Figure 3). Since pruning

Method	UCF Weizmann Dynamic
Our Baseline (Section 3)	36.5%
Automatic Localization + Interest Point Pruning	41%
Automatic Localization + Interest Point Pruning + Cuboid Masking	48%

Table 4. The above table shows the accuracy on UCF Weizmann Dynamic dataset using combination of Interest Point Pruning (IPP) and Cuboid Masking (CM) w.r.t **Automatic masks**. We can see that optimal accuracy is achieved when using both IPP and CM strategies.

Method	UCF Weizmann Dynamic
Our Baseline (Section 3)	36.5%
Ground-truth Localization + Interest Point Pruning	68%
Ground-truth Localization + Interest Point Pruning + Cuboid Masking	89%

Table 5. The above table shows the accuracy on UCF Weizmann Dynamic dataset using combination of Interest Point Pruning (IPP) and Cuboid Masking (CM) w.r.t **Ground truth masks**. We can see that optimal accuracy is achieved when using both IPP and CM strategies.

helps remove irrelevant interest points in the UCF Weizmann Dynamic dataset, we see improvement in recognition results (see Table 3). We also present the best possible recognition accuracy that can be achieved using ground-truth localization masks.

Although there is improvement in recognition accuracy for the UCF Weizmann Dynamic dataset, it is still not comparable to that achieved on the Weizmann dataset (even when using ground-truth localization). This can be attributed to the presence of background information within the cuboids extracted around the relevant interest points. This background is incorporated in the descriptor construction process and thus negatively affects performance.

In the next section, we will discuss actions that are more prone to the presence of background in extracted cuboids and how localization can be used to eliminate this irrelevant information.

5.3. Cuboid Correction

Previously, we showed how generating automatic action localization and using it to prune interest points helps improve recognition accuracy on the new synthesized complex dataset. However, the results obtained (refer to Table 3) are still not comparable to those achieved by baseline systems on Weizmann dataset. In this section, we will explore the problem further and show how eliminating background information from within relevant cuboids further improves results.

Moving actions (e.g. running, walking) are more prone to be affected by complex backgrounds than stationary actions (e.g. bending, waving). Despite pruning interest points, cuboids may still contain background pixels; cuboids extracted near the mask boundary contain irrelevant spatial information while cuboids extracted for fast moving actions (such as legs of running and walking) contain temporal background information. To deal with this, we make use of localization masks by forcing all pixels of the extracted cuboids, that lie outside the localization bounding region, to a constant value. This helps *mask* out the irrelevant background pixel values, resulting in similar gradients across same actions in the descriptor construction phase. This modification to the cuboid is what helps in optimal results on the new synthesized complex dataset.

An illustration of this is shown in Figure 4 for the UCF Weizmann Dynamic dataset. Each row shows the *same* running action performed by the *same* person on *different* dynamic backgrounds. The 2nd column shows some of the extracted cuboids of the corresponding video sequence while the 3rd column shows the same cuboids after applying cuboid masking. The 4th shows temporal gradients corresponding to column 2 while the 5th column shows temporal gradients corresponding to column 3.

For convenience, we highlight cuboid frames showing background pixels in column 2 through 5 with a red outlining. We observe that the background content in the cuboids (column 2) varies significantly for each video, leading to different temporal gradients (column 4) and eventually different descriptors. Although all 3 videos are of the same action, differences in background force systems to index these videos under different classes and thus decrease overall recognition performance.

On the contrary, application of our cuboid masking technique handles this problem. Column 3 shows how all cuboid frames composed of background content are blackened out. As a result, temporal gradients associated with background information inside cuboids (column 5) are highly similar for each of the action video. This helps in assigning the same label for all 3 videos and thus improve recognition performance.

To strengthen our case, we measure the average structural similarity (SSIM) for temporal gradients with and without cuboid masking of all 3 videos shown in Figure 4. We found the average SSIM value to be 0.67 for the case without cuboid masking and 0.75 for the case with cuboid masking. With higher SSIM score, it is evident that cuboids gradients are more similar after cuboid masking and hence improve the recognition results.

Tables 4 and 5 shows results associated with cuboid

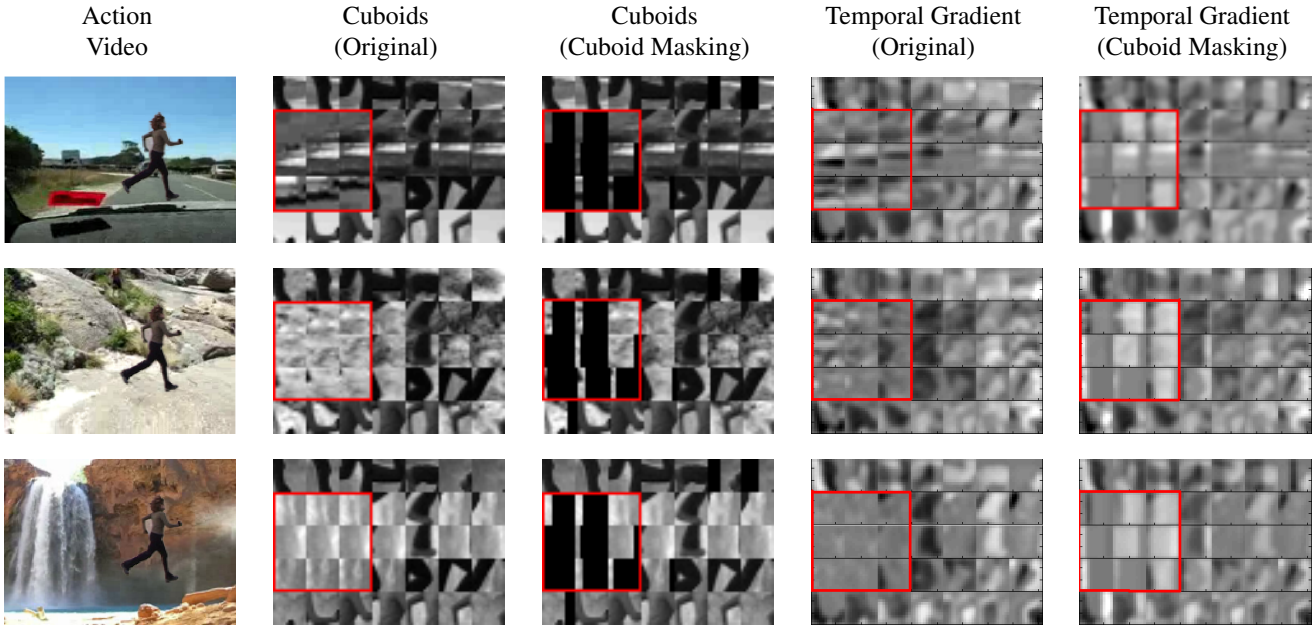

Figure 4. The figure shows the effect of cuboid masking. **Column 1:** Shows the same running action performed by the same person matted on 3 different complex moving backgrounds. **Column 2:** Shows cuboids extracted from each video sequence. Size of each cuboid is 13x13x7, where all 7 frames are shown in a single row. **Column 3:** Illustrates the exact same cuboids as in column 2 after applying cuboid masking. **Column 4:** Shows the temporal gradients of cuboids in column 2. **Column 5:** Shows the temporal gradients of cuboids in column 3. The gradient in column 4 corresponding to background content (red outlined) appear different for each video sequence. However, the gradients of all three actions looks similar after applying cuboid masking, as depicted in column 5. This is confirmed by average SSIM values of 0.67 and 0.75 for original temporal gradients (column 4) and cuboid masked temporal gradients (column 5) respectively.

masking for both automatic and ground-truth localization. We see an improvement of 11.5% and 52.5% respectively over the baseline results. We can see that even with an average automatic localization method, we are able to achieve more than 10% improvement over the baseline performance. This is a significant jump in performance and shows how cuboid masking is able to handle complex dynamic backgrounds. With better localization techniques however, there is scope of even more improvement as depicted by the results obtained using ground-truth localization.

Having analyzed the problem using the synthesized datasets, we next test our system on a realistic dataset. Instead of Youtube [17, 15] and Hollywood [12, 18] datasets, we used the UCF Sports dataset for this task. The reason for this choice being that the UCF Sports dataset is more coherent with regards to the action categories as opposed to both Youtube and Hollywood datasets.

6. UCF Sports

In order to show that our suggestions are applicable to real life datasets, we test our system on the UCF Sports datasets. UCF Sports dataset has the complex background and camera movement which were simulated in the synthetic dataset. At the same time, actions are more coherent and well captured unlike Youtube and Hollywood.

Method	UCF Sports
Our Baseline (Section 3)	68%
Automatic Localization + Interest Point Pruning	77%
Automatic Localization + Interest Point Pruning + Cuboid Masking	80%

Table 6. The table shows the results on UCF Sports with **Automatic mask**. It is evident that interest point pruning (IPP) and cuboid masking (CM) strategies improve the accuracy by 12%

The results of different experiments on this dataset are presented in tables 6 and 7. We see that automatic localization alone does not improve results but when combined with cuboid masking, we see a 12% improvement over the baseline results. We also tested using ground-truth masks for the best possible results and observed a 17% improvement over the baseline results. Using either automatic or ground-truth localization, we observe that application of localization for the purpose of interest point pruning is not sufficient. It is the use of localization to correct cuboid corruption that leads to significant improvement over the baseline method.

Method	UCF Sports
Our Baseline (Section 3)	68%
Ground-truth Localization + Interest Point Pruning	79%
Ground-truth Localization + Interest Point Pruning + Cuboid Masking	85%

Table 7. The table shows the results on UCF Sports with **Ground-truth mask**. It is evident that interest point pruning (IPP) and cuboid masking (CM) strategies improve the accuracy by 17%

7. Discussion and Conclusion

In this paper, we introduce a new synthesized, complex dataset which we argue is better suited for analyzing how recognition is affected in presence of background complexity. We show how a change from simple to complex background significantly affects the performance of traditional recognition tools. Using our new synthesized complex dataset, we establish that drop in accuracy is directly related to localization and its application in eliminating background information from the recognition pipeline. A detailed analysis of the new dataset is presented, with special emphasis on the impact of factors such as background gradients, background motion and action localization on the recognition results. In light of the analysis, we show how person localization combined with cuboid modifications helps tackle the background complexity problem and thus substantially improve overall recognition results. We show how 'proper' use of localization for interest point pruning and cuboid modification leads to a substantial increase in performance accuracy on both the synthesized and realistic datasets. An automatic localization method is also presented which is shown to outperform the baseline approach. Results are shown with ground-truth masks to show how near-perfect localization helps in improving the recognition accuracy.

Acknowledgements

This work was supported by NSF grants IIS-0905387 and IIS-0916868.

References

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *The Tenth IEEE International Conference on Computer Vision (ICCV'05)*, pages 1395–1402, 2005.

[2] M. Bregonzio, S. Gong, and T. Xiang. Recognizing action as clouds of space-time interest points. In *CVPR*, 2009.

[3] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, pages 65–72, 2005.

[4] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Discriminatively trained deformable part models, release 4. http://people.cs.uchicago.edu/ pff/latent-release4/.

[5] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1627–1645, 2010.

[6] P. V. Gehler and S. Nowozin. Let the kernel figure it out: Principled learning of pre-processing for kernel classifiers. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8, 06 2009.

[7] A. Gilbert, J. Illingworth, and R. Bowden. Fast realistic multi-action recognition using mined dense spatio-temporal features. In *IEEE 12th International Conference on Computer Vision (ICCV)*, 2009.

[8] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. In *CVPR*, pages 2376–2383. IEEE, 2010.

[9] N. Ikizler-Cinbis and S. Sclaroff. Object, scene and actions: Combining multiple features for human action recognition. In *ECCV*, 2010.

[10] Z. Jiang, Z. Lin, and L. S. Davis. A tree-based approach to integrated action localization, recognition and segmentation. In *Third Workshop on Human Motion (In Conjuntion with ECCV)*, 2010.

[11] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference*, pages 995–1004, sep 2008.

[12] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, 2008.

[13] A. Levin, D. Lischinski, and Y. Weiss. A closed-form solution to natural image matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:228–242, 2008.

[14] Z. L. Liangliang Cao and T. S. Huang. Cross-dataset action detection. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2010.

[15] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:461–468, 2009.

[16] J. Liu and M. Shah. Learning human actions via information maximization. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, 2008.

[17] J. Liu, Y. Yang, and M. Shah. Learning semantic visual vocabularies using diffusion distance. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:461–468, 2009.

[18] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. *IEEE Conf. Computer Vision and Pattern Recog*, 2009.

[19] K. Schindler and L. van Gool. Action snippets: How many frames does human action recognition require? *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1–8, 2008.

[20] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conference*, page 127, sep 2009.