Hollywood 3D: Recognizing Actions in 3D Natural Scenes

Simon Hadfield Richard Bowden
Centre for Vision, Speech and Signal Processing
University of Surrey, Guildford, Surrey, UK, GU2 7XH
{s.hadfield,r.bowden}@surrey.ac.uk

Abstract

Action recognition in unconstrained situations is a difficult task, suffering from massive intra-class variations. It is made even more challenging when complex 3D actions are projected down to the image plane, losing a great deal of information. The recent emergence of 3D data, both in broadcast content, and commercial depth sensors, provides the possibility to overcome this issue. This paper presents a new dataset, for benchmarking action recognition algorithms in natural environments, while making use of 3D information. The dataset contains around 650 video clips, across 14 classes.

In addition, two state of the art action recognition algorithms are extended to make use of the 3D data, and five new interest point detection strategies are also proposed, that extend to the 3D data. Our evaluation compares all 4 feature descriptors, using 7 different types of interest point, over a variety of threshold levels, for the Hollywood3D dataset. We make the dataset including stereo video, estimated depth maps and all code required to reproduce the benchmark results, available to the wider community.

1. Introduction

This paper presents a new 3D dataset for Action Recognition in the Wild. The detection and recognition of actions in natural settings is useful in a number of applications, including automatic video indexing and search, surveillance and assisted living. Benchmark datasets such as KTH [21] or Weizmann [2] have been invaluable in providing comparative benchmarks for competing approaches. However, high performance rates are routinely reported on these staged datasets and this suggests that they are reaching the end of their service to the community. More recent datasets such as Hollywood [14] and Hollywood2 [16] attempt to provide a more challenging problem and consist of actions "in the wild" consisting of video clips taken from a variety of Hollywood feature films. These datasets presented a new level of complexity to the recognition community, arising from the natural within-class variation of unconstrained data, including unknown camera motion, viewpoint, lighting, background and actors, and variations in action scale, duration, style and number of participants. While this natural variability is one of the strengths of the data, the lack of structure or constraints make classification an extremely challenging task.

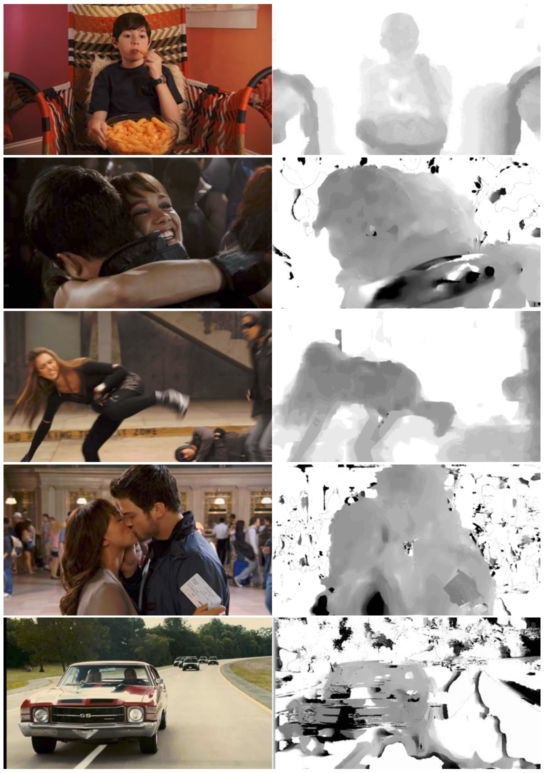


Figure 1: Example frames of various action sequences from the dataset, showing the left viewpoint and depth streams. Darker regions of the depth image are closer to the camera. From top to bottom the actions are *Eat*, *Hug*, *Kick*, *Kiss* and *Drive*.

In this work, a new natural action dataset is introduced termed Hollywood3D (see figure 1), it builds on the spirit of the existing Hollywood datasets but includes 3D information. This 3D information gives additional visual cue's which can be used to help simplify the within-class variation of actions. Lighting variations are generally not expressed in depth data, and actor appearance differences are eliminated (although differences in body shape remain). Additionally, depth provides useful cues for background segmentation, and occlusion detection.

The recent introduction of affordable 3D capture devices such as the Microsoft Kinect, has resulted in an explosion of techniques based on 3D data. Furthermore, the uptake of 3D display technology in the home and Cinema, has prompted television networks to begin broadcast of 3D programming, and film studios to produce commercially available 3D films. In this work, data is extracted from the latter commercially available sources, providing a number of advantages over self-captured data with a depth sensor[15, 5, 4]. The type of actions that you get in movies or "in the wild" is substantially different from the more contrived set-ups that exist in "lab-setting" datasets. Methods designed/trained on the latter rarely work on the former. Commercial data offers a richer range of actors, locations and lighting conditions than could be easily achieved in a lab and is one of the strengths of the Hollywood [14] and Hollywood2 [16] datasets. Additionally, active depth sensors are often unable to function in direct sunlight, severely limiting possible applications. Finally, 3D information produced by active depth sensors tends to be much lower fidelity than that available commercially, and is limited in terms of operational range. In addition to the release of a new dataset, which incorporates both original video and depth estimates, this paper provides baseline performance using both depth and appearance, and the software necessary to reproduce these results.

Previous work on action recognition, has focused on the use of feature points which can either be sampled densely or sparsely within the video. Sparse sampling reduces the impact of large background regions while dense sampling can capture context. However, all approaches sample from the spatio-temporal domain using visual appearance in x, y, and t. In this work, the additional dimension z is employed, and we show how this depth information can be incorporated both at the descriptor level, and while detecting regions of interest, extending common Spatio-temporal Interest Point techniques.

The remainder of the paper is structured as follows. The state of the art in natural 2D action recognition is first discussed in section 2, followed by section 3 covering the data extraction process, with details of the dataset. Section 4 provides a general overview of the action recognition methodology employed. Section 5 details the depth-aware spatio-temporal interest point detection schemes, followed by extensions for two state of the art feature descriptors in section 7. Results are provided with different combinations of interest point and recognition schemes in sections 8. Finally section 9 draws conclusions about the benefits of depth data in natural action tasks, and the relative merits of the presented approaches.

2. Related Work

The majority of existing approaches to action recognition focus on collections of local feature descriptors. These descriptors can be applied sparsely, i.e. at areas detected as being "interest points", or densely, using a regular sampling scheme. However, for reasons of scalability, the sparse sampling scheme is often favored. These interest points detect salient image locations, for example using separable linear filters [7] or spatio-temporal Harris corners [13]. Descriptors are generated around these interest points in a number of ways, including SIFT and SURF approaches [26, 22, 12], pixel gradients [7], Jet descriptors [21] or detection distributions and strengths [19, 9]. Focusing on interest points allows a sparse representation, for fast computation, and reduces contamination of background regions. However, in unconstrained scenarios, the presence of dynamic backgrounds or significant camera motion can lead to overwhelming numbers of uninformative detections.

These background detections can contribute in terms of context and some authors [10, 16] take advantage of this fact, modeling context directly. By performing a separate scene classification stage, combined with prior knowledge of probable action contexts (for example the "Get Out Car" action is unlikely to occur indoors) recognition rates can be improved. The approach of Wang *et al.* [25] demonstrated that dense sampling of features provides combined action and context information, and generally outperforms sparse interest points.

Generally the local features are accumulated over the sequence to form a histogram descriptor for the entire sequence, which is then classified (often using an SVM). This accumulation provides invariance to spatial and temporal translations, and changes in speed. An alternative approach sometimes used, is to consider each frame in isolation, then to classify the video based on the sequence of frames. An example of this is assigning each frame to a state in a Hidden Markov Model (HMM), then determining the most probable action for the observed sequence of states [3]. This has the advantage that it accounts for the temporal ordering of the features, but it can be difficult to determine the correct structure of the HMM. For additional information on the current state of the art, refer to [18, 17].

3. Extracting 3D Actions from Movies

With the emergence of High Definition DVD such as BluRay$^{\text{TM}}$and the introduction of 3D displays into the consumer market, there has been a sharp rise in commercially available 3D content. However, the subset that is useful for generating an action recognition dataset is still limited. A great deal of initially available 3D films were constructed from the original 2D data, via post-processing techniques such as rotoscoping. Depth data extracted from these films is less rich, lacking depth variations within objects, resembling a collection of card board cut-outs, and is fundamentally artificial, created for effect only. Additionally, films generated entirely through CGI, such as "Monsters Inc." are unlikely to provide transferable information on human actions. For this dataset, we have focused on content captured using commercial camera rigs such as James Cameron's Fusion Camera System$^{\text{TM}}$or products from 3ality Technica. These technologies produce 3D consumer content from real stereo cameras which can be used to reconstruct accurate 3D depth maps.

The dataset was compiled from 14 films [1] and is available[2]. It contains over 650 manually labeled video clips across 13 action classes, plus a further 78 clips representing the "*NoAction*". Most 3D films are too recent to have publicly available transcriptions, and subtitles alone rarely offer action cues, so automatic extraction techniques such as those employed by Marszalek *et al.* [16] are currently not possible. For this reason manual labeling was used which ensures that all examples are well segmented from the carrier movies. In addition to the action sequences, a collection of sequences containing no actions was also automatically extracted as negative data, while ensuring no overlap with positive classes.

Actions are temporally localized to the frame level, ensuring non-discriminative data at the start and end of sequences does not confuse training, and improving separation of the *NoAction* class. The data is provided from both left and right viewpoints at 1920 by 1080 resolution, at 24 frames per second. In addition, reconstructed depth is provided for all clips, at the same resolution and frame rate. Depth is reconstructed using the bilateral grid filtering approach described in [20]. If the right appearance stream is removed from the dataset, it is possible to simulate the input data that would be provided by hybrid sensors like the Kinect, albeit at a higher spatial, and lower depth resolution. Artifacts introduced by post processing are not considered, however it may be useful in future work to examine the behavior and consequences of such artifacts, with regards to action recognition.

The 14 films comprising the dataset were split between the train and test sets on a per action basis. This means each action is tested on actors and settings not seen in the training data, emphasizing generalization. As in the Hollywood and Hollywood2 datasets, some actions occur more frequently than others, and this is represented in the dataset. Table 1 lists the number of training and test clips for each action, example images from the dataset can be seen in figure 1. The sequences last an average of 2.5 seconds each, with over 650 individual 3D actions.

Action	*NoAction*	*Run*	*Punch*	*Kick*	*Shoot*	*Eat*	*Drive*	*UsePhone*	*Kiss*	*Hug*	*StandUp*	*SitDown*	*Swim*	*Dance*	Total
Train	44	38	10	11	47	11	51	21	20	9	22	14	16	45	359
Test	34	39	9	11	50	11	47	20	20	8	21	13	17	7	307

Table 1: The number of training and test sequences available for each action in the dataset, ensuring separate films are used for training and test data.

4. Action Recognition in 4D

In this paper, recognition of actions is performed in 3 stages. Firstly salient points are detected in using a range of detection schemes which incorporate the depth information, as discussed in section 5. Next, feature descriptors are extracted from these salient points, using extensions of 2 well known techniques, discussed in detail in section 7. Finally the descriptor is classified using a Support Vector Machine (SVM).

5. Interest Point Detection

The additional information present in the depth data may be exploited during interest point extraction, in order to detect more salient features, and discount irrelevant detections. The extended algorithms discussed in this section are based on the Harris Corners work by Laptev and Lindeberg [13], the Hessian points algorithm by Willems *et al.* [26] and the Separable Filters technique by Dollar *et al.* [7]. For a comparison of these interest point detection schemes, see the survey paper by Tuytelaars and Mikolajczyk [23].

4D Harris Corners

The Harris Corner [11] is a frequently used interest point detector, which was extended into the spatio-temporal domain by Laptev *et al.* [13]. The detector is based on the second-moment-matrix (ψ) of the Gaussian smoothed spatio-temporal volume (I). Interest points are detected in the spatio-temporal volume, as locations where ψ contains 3 large eigenvalues, i.e. there is strong intensity variation along 3 distinct spatio-temporal axes. To avoid eigenvalue calculation at every point, maxima are instead detected in equation 1 (where k is typically 0.001).

[1] Avatar, Pirates Of the Caribbean: On Stranger Tides, Sanctum, Drive Angry, Spy Kids: All The Time In The World, Step Up 3D, Resident Evil: Afterlife, Fright Night, My Bloody Valentine, Tron: Legacy, A Very Harold and Kumar Christmas, The Three Musketeers, Final Destination 5 and Underworld: Awakening

[2] `personal.ee.surrey.ac.uk/Personal/S.Hadfield/hollywood3d`

$$H(u,v,w) = \text{det}\left(\psi\left(u,v,w\right)\right) - k\ \text{trace}\left(\psi\left(u,v,w\right)\right)^3 \quad (1)$$

To extend the operator into 4D, the power of the trace is increased, and ψ must be expanded to a 4 by 4 matrix, incorporating the differential (I_z) along z. However, the combination of appearance and depth streams constitutes 3.5D, rather than volumetric data (i.e. measurements are not dense along the new dimension), and so gradient calculations cannot be performed directly along the z axis. Instead, the relationship between the spatio-temporal gradients of the depth stream and those of the appearance stream are exploited. Equation 2 employs the chain rule, where I_x, I_y, I_t are intensity gradients along the spatial and temporal dimensions and D_x, D_y, D_t are the gradients of the depth stream. To aid readability, the spatio-temporal location (u,v,w) is omitted.

$$I_z = \frac{I_x}{D_x} + \frac{I_y}{D_y} + \frac{I_t}{D_t} \quad (2)$$

This allows us to define ψ in equation 3, where $g(\sigma^2, \tau^2)$ is a Gaussian smoothing function, with spatial and temporal scales defined by σ and τ respectively.

$$\psi = g(\sigma^2, \tau^2) * \begin{pmatrix} I_xI_x & I_xI_y & I_xI_t & I_xI_z \\ I_xI_y & I_yI_y & I_yI_t & I_yI_z \\ I_xI_t & I_yI_t & I_tI_t & I_tI_z \\ I_xI_z & I_yI_z & I_tI_z & I_zI_z \end{pmatrix} \quad (3)$$

The set of 4D Harris interest points $F_{4D\text{-}Ha}$ is defined as the set of spatio-temporal locations within the sequence, for which H is greater than the threshold $\lambda_{4D\text{-}Ha}$. The effect of this threshold (and the threshold of each interest point detector) on recognition performance, is examined in detail in section 8.3.

$$F_{4D\text{-}Ha} = \{u,v,w | H\left(u,v,w\right) > \lambda_{4D\text{-}Ha}\} \quad (4)$$

4D Hessian Points

In [26], Willems *et al.* extended the Beaudet Saliency Measure [1] into the spatio-temporal domain. Rather than the second-moment-matrix of Laptev *et al.* they calculated the Hessian (μ) of the Gaussian smoothed spatio-temporal volume (I). The detected interest points relate to areas with strong second order intensity derivatives, including both blobs and saddles.

As in the 4D Harris scheme, gradients along z are estimated using the relationships between the depth and intensity stream gradients. This allows the 4D Hessian μ to be calculated as in equation 5. The set of interest points $F_{4D\text{-}He}$ is calculated as the set of spatio-temporal locations, for which the determinant of μ is greater than the threshold $\lambda_{4D\text{-}He}$ as in equation 6.

$$\mu = g(\sigma^2, \tau^2) * \begin{pmatrix} I_{xx} & I_{xy} & I_{xt} & I_{xz} \\ I_{xy} & I_{yy} & I_{yt} & I_{yz} \\ I_{xt} & I_{yt} & I_{tt} & I_{tz} \\ I_{xz} & I_{yz} & I_{tz} & I_{zz} \end{pmatrix} \quad (5)$$

$$F_{4D\text{-}He} = \{u,v,w | \text{det}\left(\mu\left(u,v,w\right)\right) > \lambda_{4D\text{-}He}\} \quad (6)$$

6. Interest Points in 3.5D

In part, the Harris and Hessian interest point operators are motivated by the idea that object boundary points are highly salient, and that intensity gradients relate to boundaries. However, depth data directly provides boundary information, rendering the estimation of the intensity gradient along z somewhat redundant. An alternative approach would be to employ a "3.5D" representation, using a pair of complimentary 3D spatio-temporal volumes, from the appearance and depth sequences. Equations 7 and 8 shows this approach for the Harris and Hessian operators respectively. Where θ and ϕ are equation 1 applied to the appearance and depth streams respectively, while υ and ω are the 3 by 3 Hessians. The relative weighting of the appearance and depth information, is controlled by α. This approach exploits complimentary information between the streams, to detect interest points where there are large intensity changes and/or large depth changes.

$$F_{3.5D\text{-}Ha} = \{u,v,w | \theta(u,v,w) + \alpha\phi(u,v,w) > \lambda_{3.5D\text{-}Ha}\} \quad (7)$$

$$F_{3.5D\text{-}He} = \{u,v,w | \text{det}(\upsilon) + \alpha\text{det}(\omega) > \lambda_{3.5D\text{-}He}\} \quad (8)$$

3.5D Separable Filters

A third highly successful approach to interest point detection, is the Separable Linear Filters technique of Dollar *et al.* [7]. Peaks are detected within a spatio-temporal volume, after filtering with a 2D Gaussian c in the spatial dimensions, and a quadrature pair of Gabor filters h_{ev} and h_{od} along the temporal dimension, as shown in equation 9, where L is the input sequence.

$$S(L) = (L * c * h_{ev})^2 + (L * c * h_{od})^2 \quad (9)$$

Employing the same 3.5D approach used for the Harris and Hessian detectors, leads to equation 10, where I and D are the appearance and depth streams respectively.

$$F_{3.5D\text{-}S} = \{u,v,w | S(I(u,v,w)) + \alpha S(D(u,v,w)) > \lambda_{3.5D\text{-}S}\} \quad (10)$$

7. Feature Descriptors

When attempting to recognize actions, a variety of descriptors can be extracted from the sequence. These descriptors are frequently accumulated into histograms over the spatio-temporal volume, in order to provide invariance to temporal and spatial translations. The descriptors can be based on various types of information, including appearance, motion and saliency, however depth information has rarely been utilized.

In the following sections we describe feature extraction approaches, based on the descriptors of two widely successful action recognition schemes, extended to make use of the additional information present in the Hollywood3D dataset.

7.1. Bag of Visual Words

One of the most successful feature descriptors for action recognition is that of Laptev *et al.* [14], which incorporates appearance and motion features. Descriptors are extracted only in salient regions (found through interest point detection) and are composed of a Histogram of Oriented Gradients (HOG) G, concatenated with a Histogram of Oriented Flow (HOF) F. Both histograms are computed over a small window, storing coarsely quantized image gradients, and optical flow vectors respectively. This provides a descriptor ρ of the visual appearance and local motion around the salient point at $I(u, v, w)$.

$$\rho(u,v,w) = \left(G\left(I\left(u,v,w\right)\right), F\left(I\left(u,v,w\right)\right)\right) \qquad (11)$$

When accumulating ρ over space and time, a Bag of Words (BOW) approach is employed. Clustering is performed on all ρ obtained during training, creating a codebook of distinctive descriptors. During recognition, all newly extracted descriptors are assigned to the nearest cluster center from the codebook, and the frequency of each clusters occurrences are accumulated. In this work K-Means clustering is used, with a euclidean distance function as in [14].

To extend ρ to 4D, we include a Histogram of Oriented Depth Gradients (HODG) as shown in equation 12. Thus the descriptor encapsulates local structural information, in addition to local appearance and motion. The bag of words approach is applied to this extended descriptor, as in the original scheme. Importantly, this descriptor is not dependent on the interest point detector, provided the HODG can be calculated from the depth stream D.

$$\rho(u,v,w) = (G(I(u,v,w)), F(I(u,v,w)), G(D(u,v,w))) \qquad (12)$$

7.2. RMD

The Relative Motion Descriptor (RMD) of Oshin *et al.* has also been shown to perform well in a large range of action recognition datasets, while making use of only the saliency information obtained during interest point detection. An integral volume η is created, based on the interest point detection and their strengths. The saliency content of a sub-cuboid, with origin at (u, v, w) is defined in equation 13 as $c(u, v, w)$ for a sub-cuboid of dimensions $(\hat{u}, \hat{v}, \hat{w})$. As η is an integral volume, this can be efficiently computed using a small number of lookups. The descriptor δ of the saliency distribution at a position (u, v, w) can then be formed, by performing N comparisons of the content of two randomly offset spatio-temporal sub-cuboids, with origins at $(u, v, w) + \boldsymbol{\beta}$ and $(u, v, w) + \boldsymbol{\beta}'$ as in equation 14. Note that the collections of offsets $\boldsymbol{\beta}_{0..N}$ and $\boldsymbol{\beta}'_{0..N}$ are randomly selected prior to training, and then maintained, rather than selecting new offsets for each sequence.

$$c(u,v,w) = \sum_{\boldsymbol{\gamma}=0}^{(\hat{u},\hat{v},\hat{w})} \eta([u,v,w] + \boldsymbol{\gamma}) \qquad (13)$$

$$\delta(u,v,w) = \sum_{n=0}^{N} \begin{cases} 2^n & c([u,v,w] + \boldsymbol{\beta}_n) > c([u,v,w] + \boldsymbol{\beta}'_n) \\ 0 & \text{otherwise} \end{cases} \qquad (14)$$

By extracting δ at every location in the sequence, a histogram may be constructed, which encodes the occurrences of relative saliency distributions within the sequence, without requiring appearance data or motion estimation. Increasing the number of comparisons N leads to improved descriptiveness, however the resulting histograms also become more sparse. A common alternative is to compute several δ histograms, each using different collections of random offsets $\boldsymbol{\beta}_{0..N}$ and $\boldsymbol{\beta}'_{0..N}$. The resulting histograms are then concatenated, with the result encoding more information without sparsifying the histogram. However, this comes at the cost of the independence between bins i.e. introducing some possible redundancies.

We propose extending the standard RMD described above, by storing the saliency measurements within a 4D integral hyper-volume, so as to encode the behavior of the interest point distribution across the 3D scene, rather than within the image plane. The 4D integral volume can be populated by extracting the depth measurements at each detected interest point. RMD-4D descriptors can then be extracted, using comparisons between pairs of sub-hypercuboids. The resulting histogram encodes relative distributions of saliency, both temporally, and in terms of 3D spatial location. As with the original RMD, the descriptor can be applied in conjunction with any interest point detector. As with the 4D Bag of Words approach, these features are not restricted to the extended interest point detectors described in section 5, and work equally well with standard spatio-temporal interest points, provided that a depth video is available during descriptor extraction.

8. Experimental Results

Classification was performed for all tests, with a multi-class SVM using RBF kernels. To facilitate comparisons with the Hollywood 1 and 2 datasets, the Average Precision (AP) measure was used, as explained in the PASCAL VOC [8]. The source code for the three novel interest point detection algorithms, and the two extended Action Recognition techniques is available[3], to allow reproduction of these results. RMD tests were performed with 4 binary comparisons per histogram ($N = 4$), concatenating 10 descriptor histograms. Bag of Words tests were performed with 4,000 cluster centers (as suggested in [14]), with the HoG/HoFs calculated using a block size of 3 by 3, with 8 gradient bins and 10 flow bins.

8.1. Interest Point Analysis

First we examine the benefits of including depth information during interest point detection. The standard (i.e. 3D) Bag of Words descriptor (section 7.1) is used for classification, in conjunction with the traditional spatio-temporal interest points (Separable Filters $3D - S$ and Harris Corners $3D - Ha$) are compared to the proposed depth aware schemes. The AP for each class is shown in table 2, with bold entries indicating performance greater than both standard spatio-temporal schemes.

The type of saliency measure used has a surprisingly large effect on the performance, with the average performance for the best scheme being roughly double that of the worst, even using the same feature encoding. For the standard spatio-temporal schemes, Harris points (3D-Ha) outperform separable filter points (3D-S) for all actions. This is also reflected in the depth aware schemes, and is unsurprising, as separable filters were designed primarily for computational speed. Hessian based interest points prove less informative than the extended Harris operators in both the 4D and 3.5D case. For all detectors, the 4D scheme outperforms it's standard spatio-temporal counterpart, while the 3.5D approach proves more informative than the direct 4D extension. This confirms the belief that the calculation of intensity gradients along z is redundant, and that the combination of intensity and structure gradients, is a stronger measure of saliency.

Interestingly, certain actions consistently perform better, when described by depth aware interest points. These are actions such as *Kiss*, *Hug*, *Drive* and *Run* where there is an informative foreground object, which depth aware interest points are better able to pick out. In contrast, actions such as *Swim*, *Dance* and *Shoot* are often performed against a similar depth background, or within a group of people, and the inclusion of depth in the saliency measure is less valuable. This suggests that a combination of standard spatio-temporal, and depth aware schemes, may prove valuable.

The complexity of the depth aware interest point detectors remains of the same order as their spatio-temporal counterparts (linear with respect to u, v and w). Naturally the multiplicative factor is increased however, with 3.5D techniques being roughly twice as costly, and 4D techniques taking 4 times as long.

8.2. Descriptor Analysis

Next, the use of depth information at the feature level was explored, including it's interaction with the depth aware saliency measures. These results are shown in table 3, the correct classification rate is shown in addition to the average precision, as it is more relevant for multi-class classification tasks such as video categorization. The best performing descriptor for each saliency measure, is shown in bold.

The previously noted relationship between saliency measures, appears to hold regardless of the feature descriptor used. In all cases the fast 3D-S and 3.5D-S points, performed the worst, followed by the hessian based schemes, while the extended Harris operators provided the best performance. Also following the previously noted trend, spatio-temporal interest points offer the worst performance overall, while the 3.5D scheme prove to be the most effective way to incorporate depth information.

Both types of descriptor show a consistent improvement when incorporating structural information, with increases of roughly 30% in both average precision and correct classification. This demonstrates the value of such features for recognizing actions in the wild. Overall, the Bag of Words descriptors perform somewhat better than the RMD descriptors. This is unsurprising as the RMD relies only on interest point detections, without the inclusion of any visual and motion information.

It may have been reasonable to guess, that including structural features would prove more valuable with a standard saliency measure, as the depth information had not previously been exploited. In fact the opposite proves to be true, 4D features provide more modest gains for **3D-S** and **3D-Ha** (up to 20%) than they do when combined with extended saliency measures (up to 45%). This demonstrates that depth aware saliency measures are capable of focusing computation, into regions where structural features are particularly valuable.

The complexity of the RMD-4D is greater than the standard RMD (being linear in the range of depth values, as well as in u, v and w). This is somewhat mitigated by the use of integral volumes however, meaning that runtimes are still on the order of seconds using a single CPU. In contrast the extraction of HoDG features relates to a 50% increase in cost, while still remaining linear. However the increased feature vector length does lead to and increased cost during

[3] personal.ee.surrey.ac.uk/Personal/S.Hadfield/hollywood3d

Action	**3D-S**	**3D-Ha**	**4D-He**	**4D-Ha**	**3.5D-S**	**3.5D-He**	**3.5D-Ha**
NoAction	11.4	12.1	**12.2**	**12.9**	11.4	12.0	**13.7**
Run	12.6	19.0	15.9	**22.4**	12.7	**21.8**	**27.0**
Punch	2.9	10.4	2.9	4.8	2.9	5.7	5.7
Kick	3.6	9.3	4.2	4.3	3.8	3.7	4.8
Shoot	16.2	27.9	18.9	17.2	16.2	16.2	16.6
Eat	3.6	5.0	3.6	**5.3**	3.6	**7.7**	**5.6**
Drive	15.3	24.8	**25.6**	**69.3**	15.5	**76.5**	**69.6**
UsePhone	6.5	6.8	**14.7**	**8.0**	6.5	**17.7**	**7.6**
Kiss	6.5	8.4	**8.5**	**10.0**	6.5	**9.4**	**10.2**
Hug	2.6	4.3	3.5	**4.4**	2.6	3.4	**12.1**
StandUp	6.8	10.1	7.0	7.6	6.9	9.1	9.0
SitDown	4.2	5.3	4.5	4.2	4.2	4.3	**5.6**
Swim	5.5	11.3	7.8	5.5	5.5	5.9	7.5
Dance	2.3	10.1	4.2	**10.5**	2.2	3.8	7.5
Average	7.1	12.6	9.8	**13.3**	7.1	**13.4**	**14.1**

Table 2: Average precision per class, on the 3D action dataset, for a range of interest point detectors, including simple spatio-temporal interest points, and depth aware schemes. The Bag of Visual Words feature descriptor was used. Classes are shown in bold, when depth aware interest points outperform both 3D schemes.

Descriptor	Interest Points	CC Rate	AP
RMD	3D-S	7.2%	7.2
RMD-4D	3D-S	**7.3**%	**7.4**
HoG/Hof	3D-S	7.2%	7.1
HoG/Hof/HoDG	3D-S	**7.3**%	7.2
RMD	3D-Ha	15.3%	12.2
RMD-4D	3D-Ha	15.9%	**15.0**
HoG/Hof	3D-Ha	16.2%	12.6
HoG/Hof/HoDG	3D-Ha	**19.8**%	13.2
RMD	4D-He	10.4%	11.2
RMD-4D	4D-He	**16.2**%	**12.7**
HoG/Hof	4D-He	9.4%	9.3
HoG/Hof/HoDG	4D-He	11.4%	9.8
RMD	4D-Ha	10.7%	11.5
RMD-4D	4D-Ha	10.4%	10.3
HoG/Hof	4D-Ha	14.3%	12.5
HoG/Hof/HoDG	4D-Ha	**18.5**%	**13.3**
RMD	3.5D-S	7.3%	7.3
RMD-4D	3.5D-S	**7.6**%	**7.8**
HoG/Hof	3.5D-S	7.2%	7.1
HoG/Hof/HoDG	3.5D-S	7.3%	7.4
RMD	3.5D-He	13.3%	12.2
RMD-4D	3.5D-He	17.5%	**14.3**
HoG/Hof	3.5D-He	13.6%	11.7
HoG/Hof/HoDG	3.5D-He	**19.2**%	13.4
RMD	3.5D-Ha	12.3%	11.9
RMD-4D	3.5D-Ha	17.2%	**14.4**
HoG/Hof	3.5D-Ha	17.9%	13.0
HoG/Hof/HoDG	3.5D-Ha	**21.8**%	14.1

Table 3: Correct Classification rate and Average Precision for each combination of descriptor and saliency measure. The best feature for each saliency measure is shown in bold.

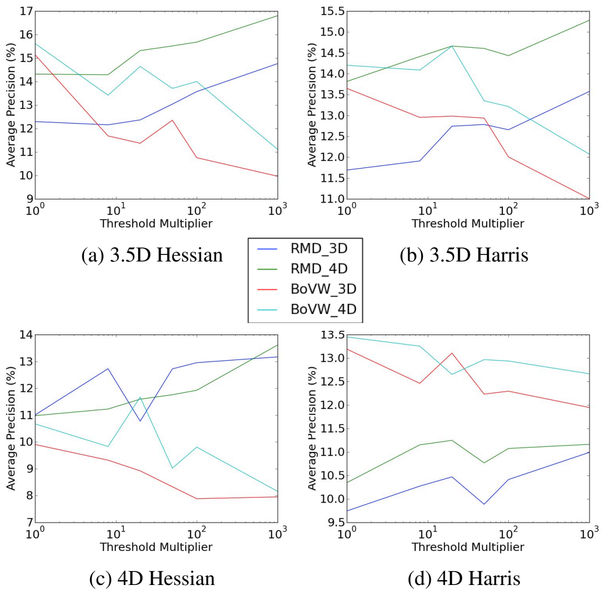
codebook generation, as k-means is generally polynomial in the number of dimensions.

8.3. Interest Point Threshold Results

Different interest point operators produce very different response strengths, meaning the optimal threshold for extracting salient points varies. In general an arbitrary threshold is selected, indeed the experiments in the previous sections employed a saliency threshold based on those suggested in previous literature. In figure 2 the relationship between the saliency threshold and the action recognition performance, is contrasted for 4D and 3.5D interest point detectors. Regardless of the saliency measure, the standard features descriptor and their depth aware extensions follow the same trend. Interestingly, for RMD based features, the trend is positive, i.e. higher saliency thresholds lead to increased accuracy. In contrast, bag of words approaches provide greater accuracy for lower saliency thresholds. This suggests that RMD features are more sensitive to noise, while the Bag of Words features are better at isolating non-discriminatory information. This makes sense, as a weak interest point relates to a single histogram entry under the bag of words scheme. In contrast, poor interest points will affect the RMD descriptor of all surrounding locations. These results are particularly interesting, as they demonstrate that interest point detection is valuable, not only for reducing the computation required, but also for improving the signal to ratio of the features. Denser features do not always lead to improved performance.

9. Conclusions

In this paper, we propose and make available a large corpus of 3D data to the community, for the comparison of action recognition techniques, in natural environments. In addition, code is also available to reproduce the baseline results presented.



(a) 3.5D Hessian

(b) 3.5D Harris

(c) 4D Hessian

(d) 4D Harris

Figure 2: Average Precision on the Hollywood 3D action recognition dataset, for various saliency thresholds, with 3.5D and 4D interest point detection.

It has been shown that 3D information provides valuable cues to improve action recognition. A variety of new interest point detection algorithm, incorporating depth data, have been shown to improve action recognition rates, doubling performance in some cases, even using standard features. Additionally 2 state of the art feature descriptors have been modified to encode structural information, demonstrating an average of 30% additional improvement in performance. Future work should focus on extending more complex featured descriptors, which may be better suited to incorporating depth information (for example Motion Boundary Histograms [6, 24]), particularly focused on mitigating the sparsity which may arise in higher dimensional feature spaces.

Acknowledgments

The work is funded under EPSRC grant EP/I011811/1 "Learning to Recognise Dynamic Visual Content from Broadcast Footage". We also wish to thank Ivan Laptev for his assistance.

References

[1] P. Beaudet. Rotationally invariant image operators. In *Joint Conference on Pattern Recognition*, 1978.

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, 2005.

[3] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *CVPR*, 1997.

[4] Chalearn. *ChaLearn Gesture Dataset (CGD2011)*, 2011.

[5] Z. Cheng, L. Qin, Y. Ye, Q. Huang, and Q. Tian. Human daily action analysis with multi-view and color-depth data. In *ECCVW*, 2012.

[6] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, 2006.

[7] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Evaluation Workshop*, 2005.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, June 2010.

[9] A. Gilbert, J. Illingworth, and R. Bowden. Fast realistic multi-action recognition using mined dense spatio-temporal features. In *ICCV*. IEEE, 2009.

[10] D. Han, L. Bo, and C. Sminchisescu. Selection and context for action recognition. In *ICCV*. IEEE, 2009.

[11] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision conference*, 1988.

[12] A. Klaser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008.

[13] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, 2003.

[14] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[15] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *CVPRW*, 2010.

[16] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2005.

[17] T. Moeslund, A. Hilton, and V. Kruger. Survey of advances in vision based motion capture & analysis. *CVIU*, 2006.

[18] O. Oshin, A. Gilbert, and R. Bowden. *Machine Learning for Human Motion Analysis*, chapter Learning to Recognise Spatio-Temporal Interest Points. IGI Publishing, 2010.

[19] O. Oshin, A. Gilbert, and R. Bowden. Capturing the relative distribution of features for action recognition. In *Face and Gesture Workshop*, 2011.

[20] C. Richardt, D. Orr, I. Davies, A. Criminisi, and N. Dodgson. Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid. In *ECCV*, 2010.

[21] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *ICPR*, 2004.

[22] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *International conference on Multimedia*, 2007.

[23] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors - a survey. *Foundations and Trends in Computer Graphics and Vision*, 2008.

[24] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011.

[25] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[26] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *ECCV*, 2008.