Modeling the Temporal Extent of Actions

Scott Satkin and Martial Hebert

Carnegie Mellon University, The Robotics Institute
{ssatkin, hebert}@ri.cmu.edu

Abstract. In this paper, we present a framework for estimating what portions of videos are most discriminative for the task of action recognition. We explore the impact of the temporal cropping of training videos on the overall accuracy of an action recognition system, and we formalize what makes a set of croppings optimal. In addition, we present an algorithm to determine the best set of croppings for a dataset, and experimentally show that our approach increases the accuracy of various state-of-the-art action recognition techniques.

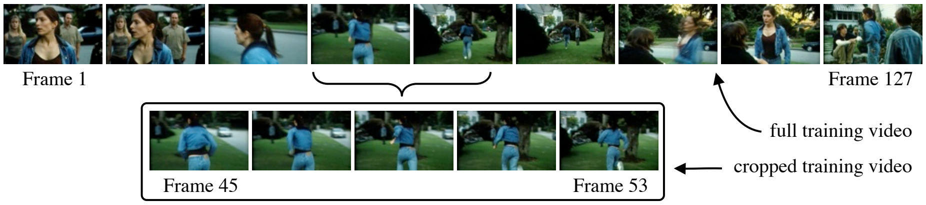

Fig. 1. The most discriminative portion of a training video is automatically extracted. The cropped training video unambiguously belongs to the action category "running" from [1].

1 Introduction

There exists an inherent ambiguity for actions – When does an action begin and end? Unlike object boundaries in static images, where one can often delineate the boundary between an object and its background, determining the temporal extent of an action is often subjective. Consider the action "eating." What is the precise moment that someone begins eating? When food is placed on a table? When a person picks up a fork? Moreover, when does the action end?

The problem is that the performance of an action recognition system may vary tremendously depending on the temporal boundaries chosen for the training samples. Researchers commonly crop training videos qualitatively based on the semantic definition of an action (which Cour *et al.* [2], Laptev *et al.* [3]

and others point out can be a very difficult task). On the contrary, we set out to automatically determine temporal croppings for videos which optimize the performance of an action recognition system. Our objective is to identify the portions of each training video in a dataset, such that if an individual video is made any shorter, it would not fully capture the true essence of the action being performed. Conversely, if a cropped video is lengthened, it would add noise to the data, making the video less discriminative.

In this paper, we formalize what makes a cropping optimal with respect to the accuracy of a trained classifier, and we present an algorithm which identifies these discriminative portions of videos. Our strategy of temporally cropping training videos is applicable no matter what representation of an action is used. Therefore, we study the effect of our method on a diverse set of action representations, and show that on a wide variety of datasets we can consistently improve the performance of a classifier by temporally cropping training videos to their most discriminative portions.

Figure 1 illustrates the concept of cropping a training video to its most discriminative portion. By running our algorithm on a video from the Hollywood-2 Human Actions and Scenes Dataset [1], we can automatically determine which portion of the video is best for detecting the "running" action. Note that unlike the original video from the dataset which contains ambiguous frames, our cropped video clearly depicts the action and disregards the frames which are not discriminative.

Collecting these types of videos, annotating their actions and delineating their boundaries is a labor-intensive task. For sufficiently large datasets, it is often impractical to do this manually. This process has been a focus of many research groups in recent years. In [1], [2], [3] and [4] the authors leverage the availability of movie scripts and closed captioning to get a rough idea of when actions occur in movies or television shows. The authors then employ various structured learning approaches to delineate these actions from their videos. Other work such as [5] and [6], focus on assisting users in the painstaking task of delineating the exact time and location of actions in videos.

Thus, it is impractical to label examples for supervised training by enforcing strict definitions of the temporal cropping of actions. Instead, our model for training involves taking video samples with approximate boundaries, and refining the samples during training. Moreover, since the temporal extent of an action is not a well-defined concept, we show that existing datasets can be further cropped during training to create a more discriminative set of training samples which improve the accuracy of a classifier, irrespective of what representation of human actions is used.

To show the broad applicability of our algorithm, we use four unique action representations: volumetric features, histograms of oriented gradients (HOG), histograms of optic flow (HOF) and point-trajectory features (Trajectons), which are a representative sampling of all major approaches. For each of these representations, we empirically show that identifying the most discriminative portions of each training video, and training a classifier on only those portions, improves overall performance.

2 Related work

Our specific problem of temporally localizing the most discriminative portions of an action can be modeled with multiple instance learning, first explored by Dietterich *et al.* [7]. Recent work has demonstrated the importance of localizing or segmenting objects from static images for the task of recognition (*e.g.*, [8], [9] and [10]). Similar methods do apply, with the key distinction that we are dealing with a single interval on the temporal axis rather than a region in the image. Buehler *et al.* [11] applied multiple instance learning in the temporal domain with the unique goal of isolating individual exemplars for actions (sign language gestures). Our effort however is focused on improving classifier performance, not finding exemplars.

Recently, there have been a few attempts to mine action recognition datasets to solve this problem. In [12], Nowozin *et al.* present an algorithm which searches for discriminative subsequence patterns in videos. However, since there is no constraint that the subsequences be continuous, this solution is equivalent to finding individual space-time features in the video which are discriminative, as opposed to our algorithm which determines the most discriminative portion of each video. Yuan *et al.* [13] propose a branch-and-bound algorithm which searches for a 3-D bounding box, akin to our temporal cropping, by maximizing mutual information of features and actions under a naïve Bayes assumption. Their method is specific to an STIP action recognition model, and cannot be applied to other systems. However, our algorithm treats the underlying action recognition system as a black box and only requires the ability to train on a subset of the dataset and evaluate its precision.

Most related to our work is that of Duchenne *et al.* [4]. Their work aims to automatically find the location of actions in videos, in a semi-supervised manner. By leveraging the availability of movie scripts and subtitles, their system begins with a rough estimate of when an action occurs. The authors then refine the location of this action using structured learning. A key distinction is that their goal is to determine temporal boundaries that approximate the way a human would qualitatively crop the data. On the contrary, our algorithm directly optimizes the accuracy of a classifier trained using the cropped videos. Unlike the authors of [4], who strive to generate croppings which perform as well as human-labeled data, we consider the performance of a classifier trained on manually cropped actions to be a baseline. Thus, we can take training data such as [4]'s “ground-truth” croppings, and further refine the temporal boundaries to produce a classifier that outperforms human-labeled data on the task of action recognition.

3 Problem formulation and overall approach

We define an "optimal set of croppings" as the set of start f^0 and end f^1 frames for each video $\mathcal{F}_i$ of class $\mathcal{C}_i$ in our training dataset which produces a classifier with the highest leave-one-out training accuracy. This can be quantified with the following high-level equation:

$$\underset{\{\forall_i:(f_i^0,f_i^1)\}}{\operatorname{argmax}} \sum_{i=1}^{n} \text{classify}\left(\text{train}(\mathcal{F}_{(1\ldots n)\neq i}, f_i^0, f_i^1),\ \mathcal{F}_i\right) = \mathcal{C}_i. \tag{1}$$

For n training videos, each with $|f|$ frames, there are $O(n^{|f|^2})$ possible sets of temporal croppings (in [4], $n = 823$ and $|f| \approx 280$). Due to this exponentially high-dimensional search space, it is intractable to test the accuracy of a classifier trained on all possible sets of croppings. Thus, a major question we address is: *How can we optimize over the massive set of potential croppings?*

In this paper we leverage the fact that portions of videos which are most confidently and correctly classified by a trained action recognition system are highly correlated with actions of the same class and differ from actions of other classes. Therefore, these portions of the videos are discriminative and are a good choice for training our classifier.

Our overall approach to determine a good cropping for an individual training video is as follows:

1. Split the video we aim to crop into its $|f|^2/2$ possible temporal croppings.
2. Train a classifier on the remaining training videos, excluding the one from step 1.
3. Evaluate this classifier on each of the $|f|^2/2$ croppings.
4. Select the individual cropping that was correctly classified with the highest level of confidence.

This approach treats the underlying action recognition system as a black box; thus, it can be applied to almost any classifier. It is a well-founded solution, which takes the form of stacked generalization [14]. Depending on the specific type of action representation being used, there are different considerations which must be taken involving tractability and the overall method of classification. Sections 4 and 5 explore two instances of this general approach: one based on space-time representations using volumetric features, the other using a more common bag-of-words representation.

4 Proof of concept experimentation

We begin by evaluating the effectiveness of our approach using Ke *et al.*'s volumetric features action recognition model [15]. This algorithm creates an action model from a single training video by segmenting a person from their background in each frame to create a 3-D silhouette. Detection is performed by comparing the boundary of this 3-D template to the edges of over-segmented frames from a testing video. We chose to experiment on Ke *et al.*'s volumetric features in this section, as a representative sample of space-time action models; although, our method can easily be applied to similar algorithms such as Rodriguez *et al.*'s "Action MACH" or Shechtman *et al.*'s "Space-Time Behavior Based Correlation" techniques [16] [17].

Since [15]'s approach builds an action recognition model from a single video, as opposed to many videos, there is only a quadratic number of croppings to consider, as opposed to the super-exponential number of possible croppings when training on multiple videos. Additionally, because the template comparison performed in [15] approximates a convolution operation, which is commutative (i.e., the template and training video can be swapped), our methodology of running a training video through a classifier as if it were a testing video to efficiently identify discriminative portions is a theoretically well-founded approximation to the high-dimensional optimization problem.

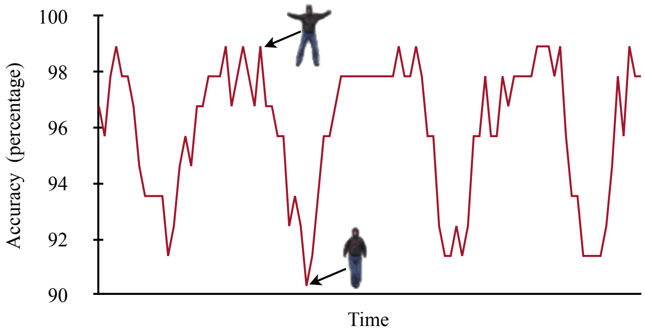

Fig. 2. Plot showing the accuracy of volumetric feature templates extracted from different times for an instance of the "jumping jack" action from the Weizmann Actions as Space-Time Shapes Dataset [18].

To quantify the benefits of temporally cropping a training video prior to creating an action recognition model, we begin by brute-force testing the accuracy of fixed length templates centered at every frame of a training video from the Weizmann Actions as Space-Time Shapes Dataset [18]. Figure 2 shows how the accuracy of a model varies based on what portion of a video it is extracted from. The periodic nature of the "jumping jack" action is quantifiable from the sinusoidal shape of the accuracy plot. A one-frame template is shown for the most discriminative and least discriminative portions of the video. It is intuitive that the most discriminative part of a jumping jack is when a person is in mid-air with all limbs extended outwards; on the contrary, when the person lands, they are momentarily indistinguishable from a person standing still. Note that the accuracies of the models vary between 90% and 99%, indicating that there is much to be gained by intelligently cropping training videos.

(a) Bend (b) Jumping Jack (c) Run (d) Wave

Fig. 3. Example categories from the Weizmann Actions as Space-Time Shapes Dataset [18].

Action	Worst Cropping Accuracy	Best Cropping Accuracy
Bend	90.63	98.00
Jumping Jack	90.94	97.70
Run	93.39	96.47
Walk	93.55	95.70
10-class Average	91.98	95.76

Table 1. Effects of cropping the Weizmann Dataset using Ke *et al.*'s volumetric feature action recognition model.

Table 1 reports the results of our approach on the entire Weizmann Actions as Space-Time Shapes Dataset [18] (shown in Figure 3) which contains 10 action classes. An interesting observation is that some classes such as "bend" and "jumping jack" have a large gap between the best and worst temporal croppings; however, actions like "run" and "walk" have less room for improvement. This is intuitive since an action like "bend" occurs at a specific instance, while "walk" has a relatively consistent appearance over time. The average accuracy of all 10 classes is also reported.

5 Temporal refinment of videos using a bag-of-words approach

Datasets such as the KTH dataset [19], and the Weizmann dataset [18] used in Section 4 have been criticized in recent years for not being a realistic sampling of actions in the real world. To tackle more complex datasets, researchers have extended the bag-of-visual-words technique from object recognition in images into the temporal domain. For example, [1], [3] and [4] represent actions as histograms of space-time interest points (STIPs), which encode both static image gradients and optic flow. The authors of [20] and [21] propose a similar method to STIPs-based systems; however, their features are based on the trajectories of tracked interest points, which can better model motion than the optic flow vectors used in previous work.

Therefore, in this section we show how our method from Section 3 can be used to determine sets of training video croppings which improve the accuracy of these bag-of-words based classifiers, which are a representative sampling of state-of-the-art techniques. We evaluate the effectiveness of our approach on two challenging datasets: Messing *et al.*'s University of Rochester Activities of Daily Living [20] and Marsazałek *et al.*'s Hollywood-2 Human Actions and Scenes Dataset [1].

5.1 Determining the best croppings

To search for the best set of croppings for each video in our training set, we augment the traditional SVM classification formulation as follows:

$$\underset{\{\forall i:(f_i^0,f_i^1)\},\mathbf{w},b,\xi}{\operatorname{argmin}} \left(\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}\xi_i \right), \tag{2}$$

$$\text{subject to:} \qquad \forall i: \quad y_i\left(\mathbf{w}\cdot\phi\left(\sum_{f=f_i^0}^{f_i^1} H_i(f)\right)+b\right) \geq 1-\xi_i. \tag{3}$$

Our max-margin formulation minimizes over f^0 and f^1, the starting and ending frames for each training video (defined in Section 3), in addition to the other standard SVM parameters. The constraints in Equation 3 include a histogram accumulation of features between start and end frames. $H_i(f)$ denotes the histogram of the quantized features from frame f of video i. As per [3] and [21], we use 4000 histogram bins for HOF and HOG space-time interest points, and 512 bins for Trajecton features. All feature vectors are $\mathcal{L}^1$ normalized prior to SVM training or classification. For consistency and simplicity, we use the same C value and a linear kernel for all experiments in this section.

Since it is infeasible to solve this high-dimensional integer linear program, we will focus on detecting the most discriminative portion of each video individually, using the approach we introduced in Section 3. This is done by training a multiclass SVM on all uncropped training videos, excluding the one which we aim to crop. We use Wu *et al.* [22]'s method of multiclass SVM classification which not only assigns category labels, but also estimates the probability that an instance belongs to each of the classes. We then evaluate the SVM on the $|f|^2/2$ possible temporal croppings of the video excluded from training to determine how discriminative each segment of the video is.

Figure 4 includes visualizations which indicate what portions of individual videos from the Hollywood-2 Dataset [4] are most discriminative. Each pixel in these images represents a different cropping (blue pixels indicate the least discriminative croppings, and red signifies the most discriminative croppings). The vertical and horizontal axis specify the starting and stopping frames (f^0 and f^1), respectively. The lower left portion of each of these figures is blank, since

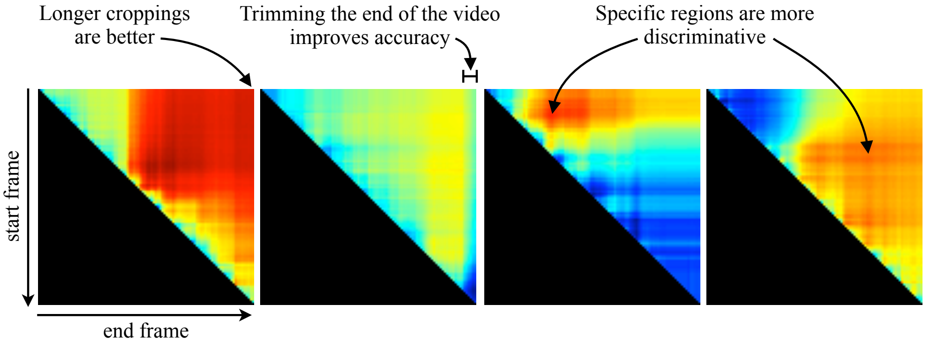

Fig. 4. Heatmaps showing which croppings are most discriminative for individual videos from the Hollywood-2 Dataset [4].

these are invalid croppings (i.e., $f^1 < f^0$). The upper right corner represents the full uncropped video. These experiments were conducted using Laptev *et al.*'s spatio-temporal feature extraction tool [3], and LIBSVM [23].

The leftmost heatmap in Figure 4 was generated from a video in which the entire video contributes to its overall discriminative quality. The farther from the diagonal $f^0 = f^1$, the longer the cropping, and the more discriminative the video gets. For videos like this one, we cannot improve the classifier accuracy with temporal cropping. On the contrary, the next heatmap shows a fascinating property inherent to many videos. The rightmost portion of the heatmap has a distinct vertical cyan stripe. This indicates that there are unusual features contained in the final frames of the video, which make the video far less discriminative. However, by trimming the last frames off of the video, independent of the starting frame, the classifier's performance can increase. Lastly, the two rightmost heatmaps in Figure 4 depict videos where most croppings would make a bad model for action recognition. For videos of this nature, it is important to choose a cropping from within the specific region we identify to be discriminative.

It is combinatorially intractable to optimize over all $O(n^{|f|^2})$ possible combinations of temporal croppings for a dataset with n training videos. Therefore, we impose the constraint: $\forall i, (f_i^1 - f_i^0)/|f_i| = \alpha$, which restricts our search space to the set of cropped video clips which are the same fixed percentage α of their full version. It is intuitive that if α is too low, we are throwing away too much of the training data; conversely, if α is unnecessarily high, we are not sufficiently cropping the training data to achieve the best possible results. Therefore, we run cross-validation to identify the ideal value of α for each dataset.

This process begins by randomly splitting the training data into two parts: a training set and a validation set. Heatmaps are generated for all videos in the training set using the leave-one-out method described above. Using these heatmaps, for a given value of α (which corresponds to a diagonal line in each heatmap), we can pick the most discriminative cropping. We then iterate over all values of α from 1% to 100%, and train a classifier on the set of croppings which corresponds to each particular value of α. The validation set of data that was withheld can now be used to determine the best value of α. By decoupling the location and length parameters of the video segments we aim to extract, we can efficiently identify discriminative combinations of cropped training videos. This approximation scheme (which requires no parameter tuning) yielded good results on all of the datasets with which we experimented.

Our algorithm scales linearly with the dataset size, and is parallelizable. Additionally, our algorithm acts as a pre-processing step which only needs to be run once prior to training. Therefore, computational expense is not a major factor.

5.2 Classification via detection

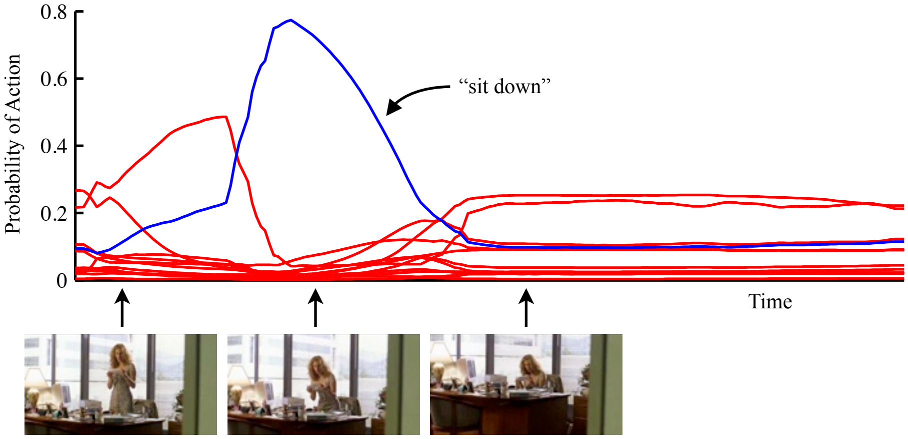

Fig. 5. The predicted probabilities for each of the 12 action classes as a function of time for a "sit down" video from [1]. The correct label is indicated in blue. Frames from the video are shown on the x-axis at their timestamp.

The standard classification paradigm of training on one set of videos, and classifying another set of videos, is not a representative problem. It is unrealistic to expect that a real-world application which uses an action recognition system would have well-cropped test data, where each video is trimmed to the length of an action. This has a major impact on how performance is evaluated (which we discuss in Section 6). Therefore, rather than extracting features from each video in its entirety, and classifying whole test samples, we employ a methodology which essentially *detects* the occurrence of an action in each video. This paradigm does not require the test videos to be cropped to the temporal extent of an individual action and can easily be altered to run on a stream of data.

For each testing video, we evaluate the multiclass SVM on a sliding window of frames. The duration of the sliding window can be set to the median length of the cropped training videos to achieve good results. We further improve the accuracy (on the order of 1%) using cross-validation to tune this parameter. During classification, we extract features from each set of frames in the sliding window, and consider each of these segments of video independently of the test video as a whole. Using [22]'s method, the probability that each of these segments belongs to individual action classes is determined.

Figure 5 shows how predicted category labels can vary drastically depending on the portion of the test video used for classification. In this example, the action "sit down" occurs almost instantaneously at approximately one-third of the way through the video (as shown in the video frames below the x-axis). For this brief portion of video, the event "sit down" (indicated in blue) is predicted by the SVM with high-confidence. At the beginning and end of the video, when the actress is either standing or has already finished sitting down, the classifier picks one of the other 12 classes (indicated in red), with a significantly lower confidence. Because we want to evaluate the performance of our algorithm in the context of a classification task, we assign a single label to an entire testing video by simply taking the peak response of the SVM classification from all timestamps.

5.3 Experimental analysis

To demonstrate the broad applicability of our approach, we evaluate the benefits of temporally cropping videos using the method described in Section 5.1 on three unique action representations: Histograms of Optic Flow (HOF) [3], Histograms of Oriented Gradients (HOG) [3] and Trajectons [21].

Our goal is to empirically show that: *By adjusting the temporal boundaries of training videos as a pre-processing step, we can improve the accuracy of a classifier, regardless of the action representation being used.* The sole purpose of the experimental analysis is to evaluate the *added benefit* of temporally cropping training videos to their most discriminative portions.

To compare our work with other action classification papers, we can only train a single multi-class SVM (and therefore can only use one set of croppings). However, our solution is general and without modification to the algorithm, we could determine a separate set of croppings for each action to train individual SVMs.

Table 2 reports the improvements from cropping the Hollywood-2 dataset prior to training. For consistency with other experiments in this paper, we use overall percentage accuracy as our performance metric. The baseline accuracy is the performance of a classifier which is trained and tested using full videos from the dataset. We compare that to the accuracy of a classifier which is trained only on the most discriminative portions of a video, using our cropping algorithm. The final columns quantify the absolute and percentage improvements due to cropping. Similarly, Table 3 reports the improvements from cropping the University of Rochester dataset prior to training. The key observation is that our strategy consistently improves the performance of an action recognition system independent of what types of features are used.

(a) (b) (c) (d)

Fig. 6. Example categories from the Hollywood-2 Human Actions and Scenes Dataset [1]. Pictured left to right: Drive, Kiss, Hug, Answer Phone.

	Baseline Accuracy (using full videos)	Our Accuracy (cropped videos)	Absolute Change cropped - full	% Improvement (cropped - full)/full
Trajectons	37.84	41.85	4.01	10.60
HOG	33.08	33.71	0.63	1.90
HOF	38.47	43.48	5.01	13.02

Table 2. Effects of cropping the Hollywood-2 Dataset using three different action representations.

(a) (b) (c) (d)

Fig. 7. Example categories from the University of Rochester Activities of Daily Living [20]. Pictured left to right: Answer Phone, Dial Phone, Drink Water, Write on Board.

	Baseline Accuracy (using full videos)	Our Accuracy (cropped videos)	Absolute Change cropped - full	% Improvement (cropped - full)/full
Trajectons	46.00	54.00	8.00	17.39
HOG	54.67	60.00	5.33	9.75
HOF	79.33	80.00	0.67	0.84

Table 3. Effects of cropping the University of Rochester Dataset using three different action representations.

6 Discussion and future work

This research has begun to explore the benefits of identifying the most discriminative portion of training videos. We presented a framework which uses a trained classifier to predict the most discriminative part of each video, irrespective of the action representation being used. Our methodology has proven to be broadly applicable, and shows the tremendous impact that the temporal cropping of videos has on the accuracy of an action recognition system. Future work will continue researching the effects of the identifying the most discriminative portions of training videos. We hope that combining multiple croppings and perhaps extending our approach to search for regions both spatially and temporally will yield even further improvement.

As a final note, while studying this problem, we noticed an important property inherent to many datasets. Because videos in most action recognition datasets are cropped to the approximate temporal extent of each action, the length of each test sample tends to be highly correlated with its action label. For example, 38% accuracy on the University of Rochester Dataset and 27% accuracy on the Hollywood-2 Dataset can be achieved by classifying solely on the number of frames in each video. This bias can easily be exploited if care is not taken to explicitly normalize for this issue. For example, it is necessary to $\mathcal{L}^1$ normalize feature histograms prior to training or classification. Not normalizing these feature vectors can lead to a substantial boost in classifier accuracy (*e.g.*, 15% increase in accuracy using Trajectons on the University of Rochester Dataset). Features themselves, such as those in [20], can also implicitly encode for the length of videos, by not limiting the number of frames which they describe.

Although using these types of features or not explicitly normalizing to ignore the number of frames in each video can yield better classification results, this is an artifact of the dataset biases and cannot be generalized to other action recognition tasks. As discussed in Section 5.2, it is not reasonable to assume that videos will be cropped tightly to the temporal extent of each action. For example, in the real-world problem of detecting the occurrence of actions in video streams, models which implicitly encode for the length of actions are no longer applicable. Moreover, if we knew how to crop these videos, this would be a solved problem. That is why we chose to implement classification as a detection problem.

This suggests ways to revisit the generation of datasets for action recognition to avoid these biases. By providing test data that is not cropped to the temporal boundaries of each action, it ensures that good action recognition systems are a result of understanding and modeling actions, not exploiting properties inherent to individual datasets.

7 Acknowledgements

This work was supported in part by the Traffic21 initiative of Carnegie Mellon University and The Technology Collaborative. We would like to thank Ivan Laptev, Pyry Matikainen and Ross Messing for sharing their datasets, source code and feature extraction utilities, and Intel Research Pittsburgh for providing computing resources.

References

1. Marszałek, M., Laptev, I., Schmid, C.: Actions in context. In: CVPR. (2009)
2. Cour, T., Jordan, C., Miltsakaki, E., Taskar, B.: Movie/script: Alignment and parsing of video and text transcription. In: ECCV. (2008)
3. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR. (2008)
4. Duchenne, O., Laptev, I., Sivic, J., Bach, F., Ponce, J.: Automatic annotation of human actions in video. In: ICCV. (2009)
5. Hall, J., Greenhill, D., Jones, G.A.: Segmenting film sequences using active surfaces. In: ICIP. (1997)
6. Wang, J., Bhat, P., Colburn, R.A., Agrawala, M., Cohen, M.F.: Interactive video cutout. ACM Transactions on Graphics **24** (2005) 585–594
7. Dietterich, T., Lathrop, R., Lozano-Perez, T.: Solving the multiple-instance problem with axis-parallel rectangles. Artificial Intelligence **89** (1997) 31–71
8. Malisiewicz, T., Efros, A.: Improving spatial support for objects via multiple segmentations. In: BMVC. (2007)
9. Uijlings, J.R.R., Smeulders, A.W.M., Scha, R.J.H.: What is the spatial extent of an object. In: CVPR. (2009)
10. Pantofaru, C., Schmid, C., Hebert, M.: Object recognition by integrating multiple image segmentations. In: ECCV. (2008)
11. Buehler, P., Zisserman, A., Everingham, M.: Learning sign language by watching TV (using weakly aligned subtitles). In: CVPR. (2009)
12. Nowozin, S., Bakir, G., Tsuda, K.: Discriminative subsequence mining for action classification. In: CVPR. (2007)
13. Yuan, J., Liu, Z., Wu, Y.: Discriminative subvolume search for efficient action detection. In: CVPR. (2009)
14. Wolpert, D.H.: Stacked generalization. Neural Networks **5** (1992) 241–259
15. Ke, Y., Sukthankar, R., Hebert, M.: Event detection in crowded videos. In: ICCV. (2007)
16. Rodriguez, M.D., Ahmed, J., Shah, M.: Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In: ICCV. (2008)
17. Shechtman, E., Irani, M.: Space-time behavior based correlation. In: CVPR. (2005)
18. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: ICCV. (2005)
19. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local SVM approach. In: ICPR. (2004)
20. Messing, R., Pal, C., Kautz, H.: Activity recognition using the velocity histories of tracked keypoints. In: ICCV. (2009)
21. Matikainen, P., Hebert, M., Sukthankar, R.: Trajectons: Action recognition through the motion analysis of tracked features. In: VOEC Workshop. (2009)
22. Wu, T.F., Lin, C.J., Weng, R.C.: Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research **5** (2004)
23. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. (2001) Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.