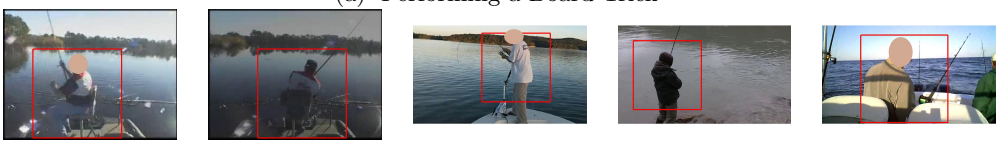
Extraction of Key Temporal Segments in Videos*

Megha Pandey, A.G.Amitha Perera, Anthony Hoogs
{megha.pandey, amitha.perera, anthony.hoogs}@kitware.com
Kitware Inc. 28 Corporate Dr., Clifton Park, NY

A video contains a large amount of visual information. In order to obtain a compact, yet distinctive representation of a video, it is important to identify the most meaningful components of the video. Shot boundary detection is often used to break down a video into smaller, semantically meaningful temporal components. In this work, our goal is to identify key temporal segments in a video clip that are most relevant to a video event category. This is achieved by discovery of recurring visual concepts across the videos from a given event class. The temporal segments of a video where this key concept appears most frequently are then labeled as the most relevant components. As opposed to some techniques of video content summarization which find most relevant objects within a single video clip, we are interested in learning the key concepts that are representative of an event class.

For automatic discovery of salient objects or visual structure in an event, we use deformable parts-based models (DPMs) with latent-SVM training [1]. A DPM is a star-structured object model which consists of a root filter, a set of part filters and a deformation model which penalizes the deviations of the part filters from their defualt locations defined with respect to the root filter. The locations of the filters are not labeled during the training and are treated as latent (hidden) variables. DPMs were used in [1] for learning object models from a set of training images annotated with ground truth object locations. [3] further extends this framework to apply it to a weakly supervised scenario where the labels for positive and negative training data are known, but the locations of the object-of-interest within each training image are not specified. This enables us to automatically learn key concepts in a set of images - without being told what these concepts are where they might be located.

(a) 'Performing a Board Trick'



(b) 'Landing a Fish'

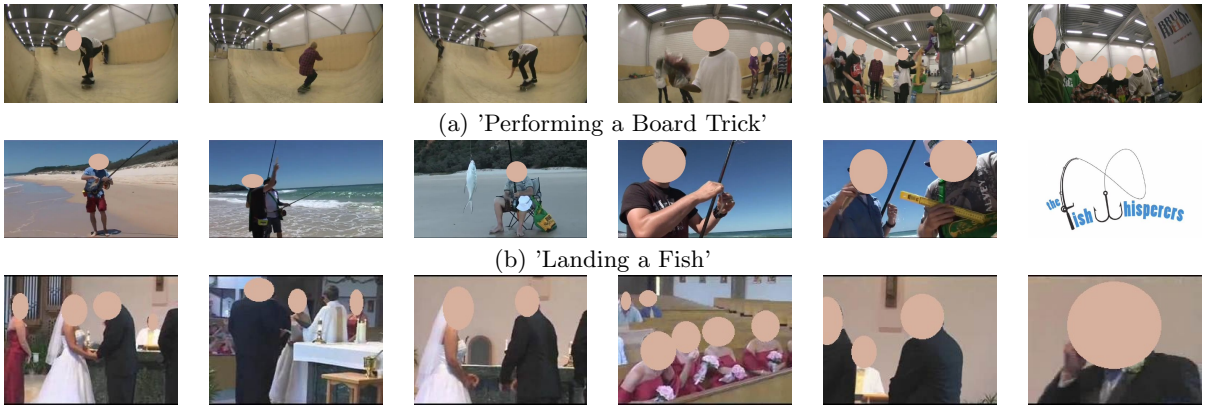
(c) 'Wedding Ceremony'

Figure 1: *Example key objects detected in different events. A bounding box marks the spatial location of the detected object.*

We apply the DPM framework as decsribed in [3] to the videos in the TRECVID MED11 dataset [2] and learn a model for each event class. Keyframes extracted from the given event videos are used as positive training images, keyframes from all the other events comprise the negative training set. The training process learns a model for an object or pattern which is common to the event-of-interest but absent from the remaining events. The spatial location of the common object discovered within a keyframe is returned as well. This location is decsribed by a bounding box and is the position where the model gets the highest score. The model so learnt can now be applied to the frames from any video to detect the presence of the corresponding key concept. A high DPM score indicates a high likelihood of the key concept being detected in that frame. Figure 1 shows some of the high scoring keyframes for different event models and gives an example of the key conepts learnt by the DPM model.

*Topic: visual processing and pattern recognition. Preference: poster.

The model scores can also be used to identify relevant temporal segments within a video. In order to do so, the DPM model is applied to the frames of a video clip from the corresponding event. The model scores are then averaged over each shot in the video clip. The shots containing several instances of the key concept that is described by the DPM model will get a high average score and are more likely to be relevant to the event. Figure 2 shows some qualitative results. Each row contains shots from one video clip, ranked from higher to lower DPM score. The middle frame of each shot is shown as a representative frame. Based on the DPM scores, the relevance of these images to the event, decreases from left to right in each row.



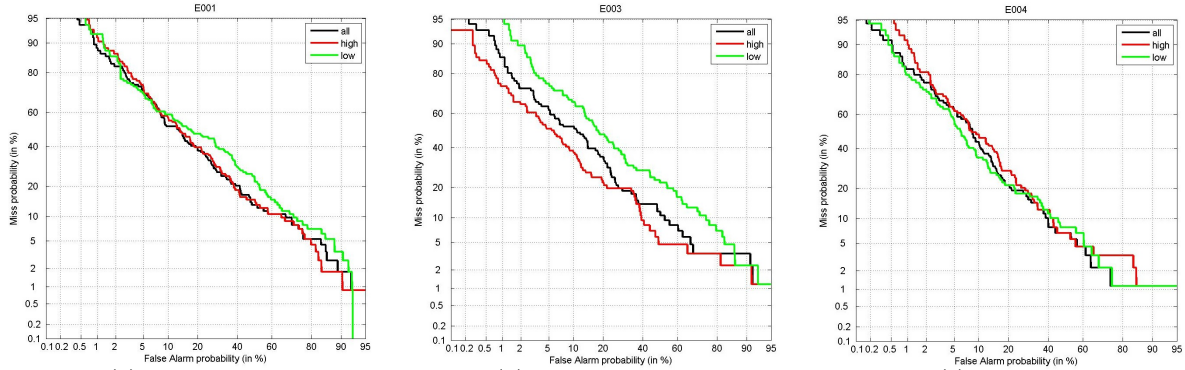
(a) 'Performing a Board Trick'

(b) 'Landing a Fish'

(c) 'Wedding Ceremony'

Figure 2: *Video segments ranked according to average DPM scores. Each row represents one video clip with the corresponding shots sorted from highest to lowest score.*

To quantitatively assess the performance of DPMs in ranking the shots of a video clip, we use these shot rankings in a video event classification task. By thresholding the DPM score, we divide all the shots from the positive training video clips into two sets. We learn an SVM model using the sets of both higher-than-threshold and lower-than-threshold shots respectively. In our experiments, HOG3D bag-of-words descriptor is used to represent each shot. A histogram intersection kernel (HIK) distance function is used to learn the SVM model. The SVM models are then used to evaluate a set of unseen videos. To compare the classification performances, we plot the results as detection-error trade-off (DET) curve. Figures 3(a) & (b) show that the classification performance is poorer when only low scoring shots are used in training. These observations confirm that the higher scoring shots retain the temporal segments most pertinent to the event, while low scoring shots are noisier. In Figure 3(c), however, the three curves are very similar. This is the due to the fact that the video clips from *Wedding* event do not have as much temporal variation and thus the high scoring and low scoring shots are largely similar.



(a) Board Trick (b) Landing a Fish (c) Wedding

Figure 3: *DET curves. Classification performance using shots with higher (resp. lower) than threshold DPM scores is plotted in red (resp. green). Classification performance when using the entire video clip is shown in black.*

References

[1] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 2010.

[2] Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Wessel Kraaij, and Alan F. Smeaton. Trecvid 2011 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2011*. NIST, USA, 2011.

[3] Megha Pandey and Svetlana Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, 2011.

Acknowledgements

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20069. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.