A Multi-scale Boosted Detector for Efficient and Robust Gesture Recognition

Camille Monnier(✉), Stan German, and Andrey Ost

Charles River Analytics, Cambridge, MA, USA
cmonnier@cra.com

Abstract. We present an approach to detecting and recognizing gestures in a stream of multi-modal data. Our approach combines a sliding-window gesture detector with features drawn from skeleton data, color imagery, and depth data produced by a first-generation Kinect sensor. The detector consists of a set of one-versus-all boosted classifiers, each tuned to a specific gesture. Features are extracted at multiple temporal scales, and include descriptive statistics of normalized skeleton joint positions, angles, and velocities, as well as image-based hand descriptors. The full set of gesture detectors may be trained in under two hours on a single machine, and is extremely efficient at runtime, operating at 1700fps using only skeletal data, or at 100fps using fused skeleton and image features. Our method achieved a Jaccard Index score of 0.834 on the ChaLearn-2014 Gesture Recognition Test dataset, and was ranked 2nd overall in the competition.

Keywords: Gesture recognition · Boosting methods · One-vs-all · Multi-modal fusion · Feature pooling

1 Introduction

Automated gesture recognition has many desirable applications, including home entertainment, American Sign Language (ASL) translation, human-robot interaction (HRI), and security and surveillance. The area has been a focus of extensive research and development in the past decade, and while significant advances in sensor technologies and algorithmic methods have been made, current systems remain far from capable of human-level recognition accuracy for most real-world applications. The problem of recognizing gestures in a stream of data comprises multiple challenges, including noisy or missing data, non-uniform temporal variation in gesture execution, variability across individuals, and significant volumes of data. In this paper, we present an efficient and highly-competitive approach to detecting and recognizing gestures in a stream of multi-modal (skeleton pose, color, and depth) data. Our proposed method achieved a Jaccard Index score of 0.834 on the ChaLearn-2014 Gesture Recognition Test dataset, and was ranked 2nd overall in the competition.

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part I, LNCS 8925, pp. 491–502, 2015.
DOI: 10.1007/978-3-319-16178-5_34

1.1 Related Work

Vision-based pose and gesture recognition technologies have been developed for three general categories of sensors: monocular cameras, stereo cameras, and fused color/active ranging sensors such as the Microsoft Kinect. Monocular (single-camera) methods offer an advantage in terms of cost and flexibility of application, but represent a significant challenge from an algorithmic perspective as depth, segmentation and pose data are not easily obtained. Much of the foundational work in pose and gesture recognition addresses monocular imagery [1,2], and this continues to be an active area of research [3–5]. Stereo cameras, which provide depth information in addition to color imagery and function equally indoors and outdoors, have recently been applied to the problem of recognizing gestures in real-world applications such as gesture recognition for robot control [6,7].

The Kinect series of sensors, which provides built-in skeleton tracking data along with high-resolution depth and co-registered color imagery, has been applied to a wide variety of tasks involving gesture recognition, including entertainment, human-robot control [8], and virtual telepresence [9]. While the technology is restricted to indoor use, the rich data produced by the Kinect is particularly well-suited to the task of recognizing complex human gestures such as american sign language (ASL), as well as culturally significant gestures and body language.

Researchers have developed a variety of approaches to extracting features suitable for representing complex gestures or gesture elements, including bag-of-words representations [10], poselets [11,12], and hierarchical representations [13,14]. Significant effort has been aimed at developing methods capable of discriminating between complex temporal sequences. Popular models reported in the gesture recognition literature typically derive from sequence-learning methods such as Hidden Markov Models (HMM) and Conditional Random Fields (CRF) [15]. Song *et al.*[13] propose a hierarchical sequence summarization (HSS) approach to recognizing gestures based on CRFs. The winners of the 2013 ChaLearn competition combine an HMM audio classifier with a dynamic time warping (DTW) based pose sequence classifier [16].

So-called "non-temporal" models that operate on fixed-length sequences, such as Support Vector Machines (SVM) [17], Random Decision Forests (RDF) [18], and boosting methods [19] have been successfully applied to the problem of gesture and action recognition[14,20], but are often passed over in favor of models that are expected to implicitly handle complex temporal structures, such as HMM and CRF [13,21]. In this paper, we demonstrate that non-temporal methods such as Adaboost can indeed yield highly-competitive results for gesture detection when combined with appropriate multi-scale feature representations.

2 Proposed Method

We propose an approach to gesture recognition that combines a sliding-window detector with multi-modal features drawn from skeleton, color, and depth data produced by a first-generation Kinect sensor. The gesture detector consists of a set of boosted classifiers, each tuned to a specific gesture. Each classifier is

trained independently on labeled training data, employing bootstrapping to collect hard examples. At run-time, the gesture classifiers are evaluated in a one-vs-all manner across a sliding window. Fig. 1 illustrates our multi-scale approach to feature extraction and gesture classification. We describe the dataset, features, and classifiers in the following sections.

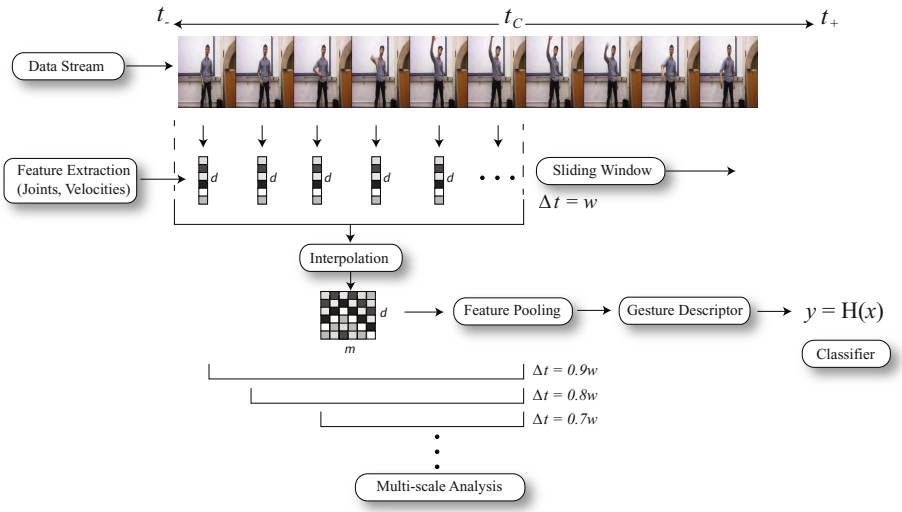

Fig. 1. Overview of our sliding-window approach. Local features are extracted on a frame-by-frame basis, and interpolated to a fixed-size feature matrix. Following a pooling step, the resulting descriptor is processed by a previously trained classifier. The process is repeated at multiple scales and offsets in the data stream.

Table 1. ChaLearn 2014 Gesture Dataset Statistics

Dataset	Labeled Instances	Length (min)
Development	7,754	∼470
Validation	3,362	∼230
Test	2,742	∼240

3 Dataset

We report methods and results developed on the ChaLearn 2014 Gesture Recognition challenge dataset. The challenge dataset consists of *Development*, *Validation*, and *Testing* sets used throughout different stages of the competition. Each dataset respectively contains 470, 230, and 240 minute-long sequences of culturally-relevant gestures performed by multiple individuals. Fig. 1 describes

statistics for each dataset. The challenge focuses on a specific set of 20 labeled Italian gestures, but includes multiple unlabeled gestures as confusers. Data products include color and depth video, segmentation masks, and skeleton joint data produced by a first-generation Kinect sensor. Fig. 2 illustrates sample data for a single gesture sequence.

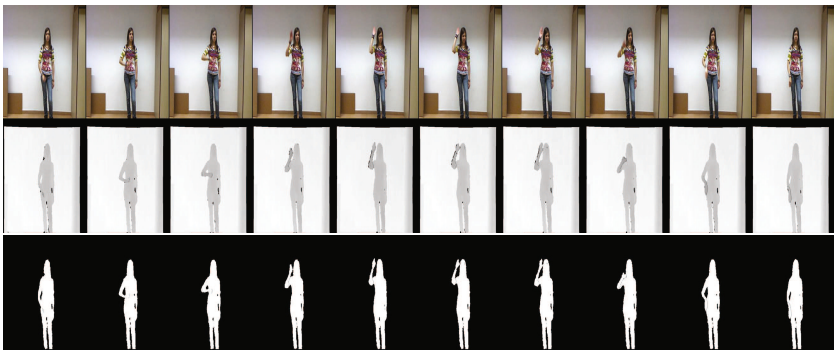

Fig. 2. Example color, depth, and segmentation data corresponding to a single gesture instance

4 Features

We extract features including normalized skeleton joint positions, rotations, and velocities, as well as HOG descriptors of the hands. Features are extracted at multiple temporal scales to enable recognition of variable-length gestures. Features including skeleton pose and hand shape are extracted from each frame of video to produce a corresponding sequence of d-dimensional descriptors. These descriptors are interpolated to produce a fixed-width $m \times d$ sequence describing the sequence, where m is the expected minimum duration, in frames, of a single gesture. Features in this sequence are then pooled to produce a final descriptor that may be processed by a gesture classifier. This process is repeated at multiple time scales, to account for temporal variation between gesture types and across individuals.

4.1 Skeleton Features

We extract multiple features from the skeleton data produced by the Kinect, including the normalized positions of the 9 major joints in the upper body, relative positions of these joints, joint angles, and instantaneous derivatives (velocities) of each feature. To reduce variability across subjects, joint positions x_j^i for each subject i are first normalized according to the length of an individual's torso s_i, following

$$x_{i,j} = \frac{x_{i,j} - x_{i,hip}}{s_i} \tag{1}$$

Where s_i is measured as the distance between the hips and the base of the neck. Following normalization, four types of features are extracted: normalized joint positions x_j; joint quaternion angles q_j; Euclidean distances between specific joints; and directed distances between pairs of joints, based on the features proposed by Yao *et al.* [22]. This latter feature consists of the distance from one joint to another along the normal vector defined by the source joint and the its parent. The features corresponding to joint positions and angles account for $9 \times 3 \times 2 = 54$ dimensions. We compute joint-pair features between all joints with a parent node (8 joints, excluding the hip), yielding an additional $8 \times 8 = 64$ dimensions, for a total of 110 static pose features. Finally, first-order derivatives are computed across adjacent frames, for all skeletal features, bringing the total feature vector size to 220 dimensions describing the skeleton's pose and instantaneous motion in a single frame of data.

4.2 Hand Features

While many of the gestures contained in the ChaLearn-2014 dataset may be differentiated by examining the positions and movements of large joints such as the elbows and wrists, a number of gestures differ primarily in hand pose, as well as in slight differences in positioning relative to the body or face. Fig. 3 illustrates a typical set of similar gesture pairs. The first-generation Kinect provides tracking data for large joints, but does not provide tracking information for the fingers necessary to differentiate between gestures such as these.

Fig. 3. Examples of gestures that differ primarily in hand pose

We employ a straightforward approach to describing hand shape. First, a square image chip is extracted around each hand, using the position information provided by the Kinect. As the scale of the subject's hand in the image is unknown, we estimate the dimensions of each hand chip based on the known scale of a body part that is parallel to the image plane. For simplicity, we again use the torso as a reference, as the gestures are performed by upright subjects who are typically far enough from the camera for perspective effects to be negligible.

Explicitly, image dimensions for a subject's hand are computed as:

$$w_{i,h} = \frac{\|x_{i,wrist} - x_{i,elbow}\|}{s_i} s'_i \tag{2}$$

where s'_i is the length of the subject's torso as measured in image space. This approach produces an image chip scaled to the length of the subject's forearm, which is sufficient for capturing a fully-extended hand. Extracted hand images are then rescaled to 64x64 using bilinear interpolation.

To reduce the inclusion of background in the hand shape descriptor, we conduct an additional masking step using the associated depth image. Depth data is extracted using the same approach as for color images, producing a 64x64 depth image for each hand. Foreground masks are then computed by eliminating pixels whose depth deviates by more than a threshold T_d from the median depth of the image. In our experiments, T_d was computed empirically as the mean extent of well-segmented hands (i.e., hands held away from the body) in the dataset. Fig. 4 illustrates the process for producing masked hand images.

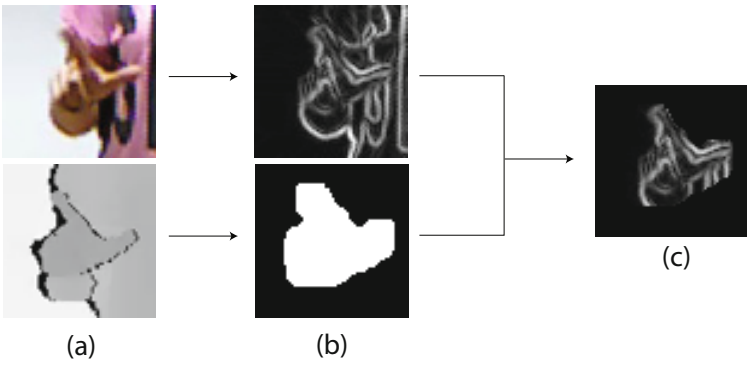

Fig. 4. Hand segmentation process. Color and depth images centered around the hand are extracted (a) using known skeleton joint positions. Depth is smoothed, thresholded to remove background, and expanded using dilation to produce a segmentation mask, and a gradient image is computed from the color image (b). The mask is applied to the gradient image (c), from which HOG features are extracted.

We compute a masked histogram of oriented gradients (HOG) descriptor [23] for each hand, using the extracted color images and depth masks. HOG features are computed for 9 orientation bins across 16x16 non-overlapping blocks, resulting in shape descriptors of dimensionality $d_{HOG} = 144$ for each hand.

4.3 Feature Pooling

Following the extraction of skeleton and hand features at each frame, features within the time window to be classified are collected and linearly interpolated

to a fixed-length sequence of size $m \times d$. To reduce sensitivity to translation and minimize noise, we perform mean pooling at multiple overlapping intervals of varying length. To capture high-level information related to gesture periodicity or complexity, we compute the variance of each feature within the same intervals used for mean pooling. Fig. 5 illustrates the pooling process. The pooled features are then combined into the final feature vector used in classification, resulting in a feature vector of dimensionality $d = 20746$.

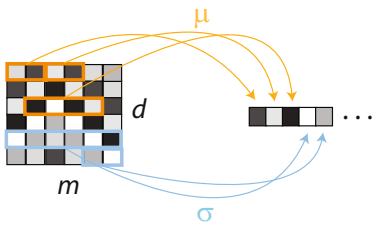

Fig. 5. Notional illustration of feature pooling. The raw feature vector is oversampled using multiple overlapping regions in which mean and variance are computed. Mean pooling is achieved using intervals and step sizes of length 2,4, and 8. Variance is computed over intervals of length 6,12 and a step size of 4. In our experiments, we define $m = 24$.

5 Classification

As sliding-window methods must typically analyze many windows at various scales, we apply an efficient boosted classifier cascade [24] to the task of recognizing individual gestures. This type of classifier provides a significant advantage in run-time efficiency over "monolithic" methods such as nonlinear SVMs, because the cascade structure enables partial evaluation (and thereby partial feature extraction) on the majority of negative samples. Each gesture is independently learned by a single boosted classifier using a one-versus-all (OVA) approach [25]. We use boosted classifiers comprising 1000 depth-2 decision trees in our experiments. During training, each classifier is initialized using the full set of labeled positive gesture examples, along with a subset of randomly sampled negative gestures and non-gestural examples. The initial gesture detector is then applied to the training data to collect hard examples, and the classifier subsequently retrained, in a process commonly referred to as bootstrapping [26,27]. The full set of gesture classifiers require approximately 2 hours to train on modern laptop, with 4 rounds of bootstrapping.

At runtime, the set of classifiers is applied to each time window, and the maximum response stored. As many overlapping windows and time scales are considered, multiple detections are typically produced in the vicinity of gestures or gesture-like sequences. To resolve conflicting detections, we apply non-maximal suppression (NMS) using the PASCAL overlap criterion [28].

6 Experimental Results

In this section, we discuss the performance of variations in our proposed method within the context of the ChaLearn competition, and provide a more detailed analysis of the top-scoring method using standard measures of detector performance.

The ChaLearn competition evaluated methods using the Jaccard Index $J \in \{0, 1\}$, which measures detection accuracy as the fractional overlap between a detection window and ground truth. Overall performance is summarized using the mean of the Jaccard Index for all truthed gesture instances. False positives are included in this statistic, and contribute a Jaccard Index value of $J = 0$.

We evaluated the performance of our proposed method using four feature sets, progressively combining: normalized skeleton joints and velocities (SK); joint angles and angular velocities (JA); joint-pair distances and velocities (JP); and hand HOG descriptors (HH). Classifiers were trained on the *Development* data, and evaluated on the reserved *Validation* data. To ensure a fair comparison across feature sets, classifier thresholds were chosen to achieve a constant rate of 1 false positive per minute (fppm). In all cases, the system was evaluated over windows computed at 30 scales, using a step size of 2 frames. Table 2 illustrates the Jaccard Index score for each variant. The baseline feature set (SK) yields a competitive score of 0.742, which is improved slightly by the inclusion of joint angle data (SK+JA). A more significant improvement is apparent from the inclusion of joint-pair features (SK+JA+JP), which likely reflects the importance of fine interactions between the various moving parts of the body, including the face and hands.

Table 2. Jaccard Index scores for detectors trained on the four feature sets. Classifiers were trained on *Development* and evaluated on *Validation* datasets

Feature set	JI Score
SK	0.742
SK+JA	0.755
SK+JA+JP	0.791
SK+JA+JP+HH	0.822

An analysis of the confusion matrix for the skeleton-only detector reveals that skeleton data is sufficient to accurately differentiate between the majority of the labeled gestures in the ChaLearn dataset, and is even sufficient to discriminate between most instances of visually similar gestures such as those illustrated in 3. The addition of hand-specific descriptors (SK+JA+JP+HH) significantly reduces error rates on these gestures, and yields our strongest detector with $J = 0.834$ on the *Test* dataset. Despite the introduction of additional unlabeled gestures in the *Test* dataset, the detector achieved higher accuracy than on the *Validation* dataset; this may be explained by the fact that the final detector was

trained on a larger dataset consisting of both *Development* and *Validation* data, and may therefore be expected to exhibit better generalization properties. The full 20-gesture detector is highly efficient, exceeding 100fps on the ChaLearn data on a single-core modern laptop. Using skeleton features only (SK+JA+JP), our detector is capable of processing over a minute of data per second, equivalent to 1700fps.

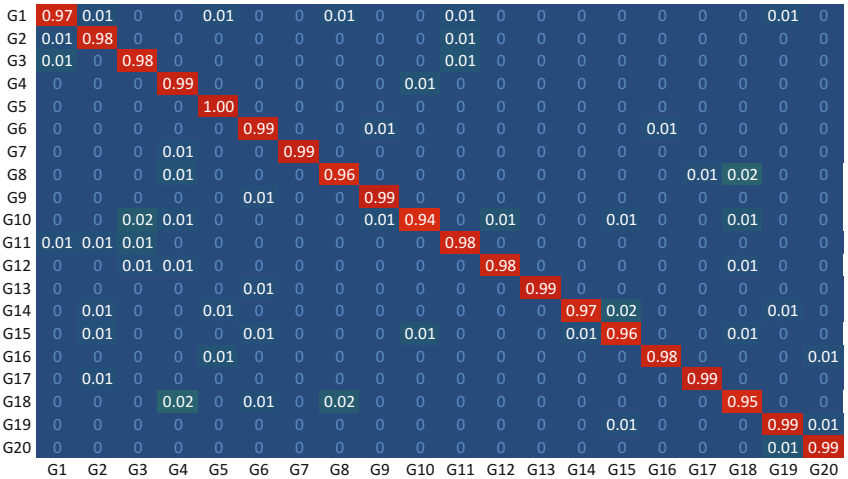

Fig. 6. Confusion matrix for the 20 hand labeled gestures in the ChaLearn-2014 *Test* dataset, produced by the SK+JA+JP+HH detector. Results are computed for detections that overlap ground truth according to the PASCAL criterion

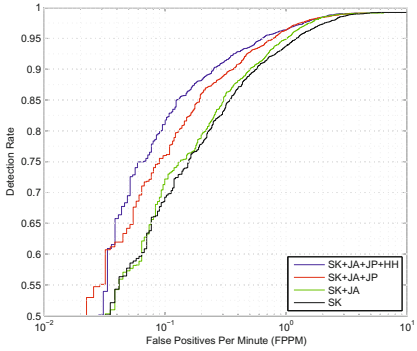

Fig. 7. Detection/false-positive tradeoff (ROC) curves for the four detectors. Although detection accuracy converges near 1 fppm, measurable differences in Jaccard Index are apparent. This difference is likely explained by the use of the PASCAL criterion in computing ROC curves, which require only 50% overlap to be considered a match.

Figure 6 illustrates the confusion matrix for the final detector on the *Test* data, computed at a 1 fppm. Recognition rates across gestures is generally consistent - detected gestures are classified with mean accuracy $97.9 \pm 1.61\%$, with a single gesture (G10) falling below 95% recognition rate. The remaining source of error in our experiments is generally attributed to false positives caused by unlabeled confusers in the *Test* data. Figure 7 illustrates the tradeoff between detection and false-positive rates.

7 Conclusions

Our approach to gesture detection achieves highly competitive results on the ChaLearn 2014 gesture recognition dataset, ranking 2nd in the overall competition. The proposed method deviates from many recently developed gesture recognition systems in its use of a boosted classifier cascade rather than sequence-learning methods such as HMM and CRF. A message from the outcome of this work is a reminder that simple methods based on effective feature construction will frequently outperform more sophisticated models that incorporate inadequate feature data. While our approach performed well on the ChaLearn dataset, it is likely that other types of gestures, such as ASL, will provide more complex structures that will pose a more significant challenge. In future work, we plan to evaluate our approach on a wider set of gesture lexicons and application areas, which may highlight specific areas for improvement.

References

1. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing **28**(6), 976–990 (2010)
2. Andriluka, M., Roth, S., Schiele, B.: Monocular 3d pose estimation and tracking by detection. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 623–630 (2010)
3. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision **103**(1), 60–79 (2013)
4. Cherian, A., Mairal, J., Alahari, K., Schmid, C., et al.: Mixing body-part sequences for human pose estimation. In: CVPR 2014-IEEE Conference on Computer Vision & Pattern Recognition (2014)
5. Yu, T.H., Kim, T.K., Cipolla, R.: Unconstrained monocular 3d human pose estimation by action detection and cross-modality regression forest. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3642–3649. IEEE (2013)
6. Nickel, K., Stiefelhagen, R.: Visual recognition of pointing gestures for human-robot interaction. Image and Vision Computing **25**(12), 1875–1884 (2007)
7. Song, Y., Demirdjian, D., Davis, R.: Continuous body and hand gesture recognition for natural human-computer interaction. ACM Transactions on Interactive Intelligent Systems (TiiS) 2(1), 5 (2012)

8. Van den Bergh, M., Carton, D., De Nijs, R., Mitsou, N., Landsiedel, C., Kuehnlenz, K., Wollherr, D., Van Gool, L., Buss, M.: Real-time 3d hand gesture interaction with a robot for understanding directions from humans. In: RO-MAN, 2011 IEEE, pp. 357–362. IEEE (2011)
9. Zhang, Z.: Microsoft kinect sensor and its effect. IEEE MultiMedia **19**(2), 4–10 (2012)
10. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8. IEEE (2008)
11. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1365–1372. IEEE (2009)
12. Holt, B., Ong, E.J., Cooper, H., Bowden, R.: Putting the pieces together: Connected poselets for human pose estimation. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 1196–1201. IEEE (2011)
13. Song, Y., Morency, L.P., Davis, R.: Action recognition by hierarchical sequence summarization. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3562–3569. IEEE (2013)
14. Wang, J., Chen, Z., Wu, Y.: Action recognition with multiscale spatio-temporal contexts. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3185–3192. IEEE (2011)
15. Mitra, S., Acharya, T.: Gesture recognition: A survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews **37**(3), 311–324 (2007)
16. Wu, J., Cheng, J., Zhao, C., Lu, H.: Fusing multi-modal features for gesture recognition. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, pp. 453–460. ACM (2013)
17. Vapnik, V.N., Vapnik, V.: Statistical learning theory. vol. 2. Wiley, New York (1998)
18. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
19. Freund, Y., Schapire, R., Abe, N.: A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence **14**(771–780), 1612 (1999)
20. Gall, J., Yao, A., Razavi, N., Van Gool, L., Lempitsky, V.: Hough forests for object detection, tracking, and action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(11), 2188–2202 (2011)
21. Tang, K., Fei-Fei, L., Koller, D.: Learning latent temporal structure for complex event detection. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1250–1257. IEEE (2012)
22. Yao, A., Gall, J., Fanelli, G., Van Gool, L.J.: Does human action recognition benefit from pose estimation?. In: BMVC, vol. 3, p. 6 (2011)
23. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)
24. Bourdev, L., Brandt, J.: Robust object detection via soft cascade. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 2, pp. 236–243. IEEE (2005)
25. Rifkin, R., Klautau, A.: In defense of one-vs-all classification. The Journal of Machine Learning Research **5**, 101–141 (2004)

26. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(4), 743–761 (2012)
27. Walk, S., Majer, N., Schindler, K., Schiele, B.: New features and insights for pedestrian detection. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1030–1037. IEEE (2010)
28. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision **88**(2), 303–338 (2010)