Entity centric Feature Pooling for Complex Event Detection

Ishani Chakraborty
SRI International
201 Washington Road
Princeton, NJ
ishani.chakraborty@sri.com

Hui Cheng
SRI International
201 Washington Road
Princeton, NJ
hui.cheng@sri.com

Omar Javed
ClipMine Inc.
520 San Antonio Rd.
Mountain View, CA
omar.javed@gmail.com

ABSTRACT

In this paper, we propose an entity centric region of interest detection and visual-semantic pooling scheme for complex event detection in YouTube-like videos. Our method is based on the hypothesis that many YouTube-like videos involve people interacting with each other and objects in their vicinity. Based on this hypothesis, we first discover an Area of Interest (AoI) map in image keyframes and then use the AoI map for localized pooling of features. The AoI map is derived from image based saliency cues weighted by the actionable space of the person involved in the event. We extract the actionable space of the person based on human position and gaze based attention allocated per region. Based on the AoI map, we divide the image into disparate regions, pool features separately from each region and finally combine them into a single image signature. To this end, we show that our proposed semantically pooled image signature contains discriminative information that detects visual events favorably as compared to state of the art approaches.

1. INTRODUCTION

Visual event detection is the problem of categorizing videos into certain pre-defined events. According to NIST [2], an event is "a happening that involves people engaged in process-driven actions with other objects". In this paper, we consider events captured in user generated, Youtube-like web videos. We are particularly interested in events involving people interacting with objects in their vicinity, for example, a person "blowing off candles" in a birthday party or a tutorial style video showing how to "change a vehicle tire". Figure 1 illustrates shots from some of these events.

Event detection in videos is a challenging problem owing to complex visual representation of participating objects as well as the semantic gap between event description and the visual components. Bags of features (BoF) strategies that represent video keyframes as orderless vectors of local feature occurrences have shown remarkable success in this domain [13]. This is mostly attributed to the spatial pooling step, which is agnostic to the spatial layout of the image and combines features from the whole image into a single BoF vector. However, recent advances suggest that augmenting

HuEvent'14, November 7, 2014, Orlando, FL, USA.

http://dx.doi.org/10.1145/2660505.2660506 .

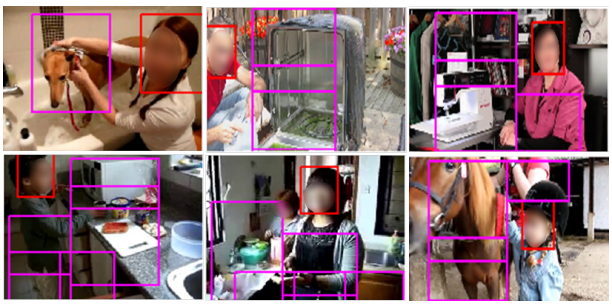

Figure 1: Video keyframes showing person-object interaction in events. The human centric actionable space is marked in magenta.

BoF with spatially *constrained* feature pooling can improve detection and reduce semantic gap. A pre-dominant approach in this direction is the spatial pyramid model [9] (SPM) that performs pooling over spatial grids or horizontal bands and combines them by concatenation of the individual BoF vectors. An alternative technique of randomized spatial partition presented in [7] shows improved performance over SPM. In it, images are randomly partitioned multiple times to obtain several independent patterns followed by retroactive selection of best partition pattern. These research endeavours clearly show that the *region of interest selection* for feature pooling strongly influences the overall performance of bag of features.

In contrast to spatial pooling from uniform or randomized regions of interest, semantics driven pooling depend on pre-trained object detectors to spatially localize individual objects [4, 6]. In [6], task-dependent objects such as water, cup and bread are detected in order to recognize the corresponding kitchen events. In Detection Bank [4], the image is first processed with a large number of object detectors and then the statistics of co-occurring objects are pooled into a spatial pyramid.

Recently, studies have shown that partitioning an image into object and non-object regions and pooling from each channel separately can improve classification accuracy of objects [14, 11]. These methods are similar to SPM, where a single image signature is generated by concatenating individual BoFs per region. However, the pooling is performed on image based object and background regions rather than pre-defined or randomized partitions. The study by Uijlings et al. [14] showed that by knowing the ideal bounding boxes of objects in images, the accuracy of object detection can be greatly increased. In general, objects are key to understanding an event and hence *searching* for relevant objects could enhance detectability of a video. However, object categorization is itself a difficult task, especially in a cluttered environment. Second, objects need to be prioritized according to their importance to the event.

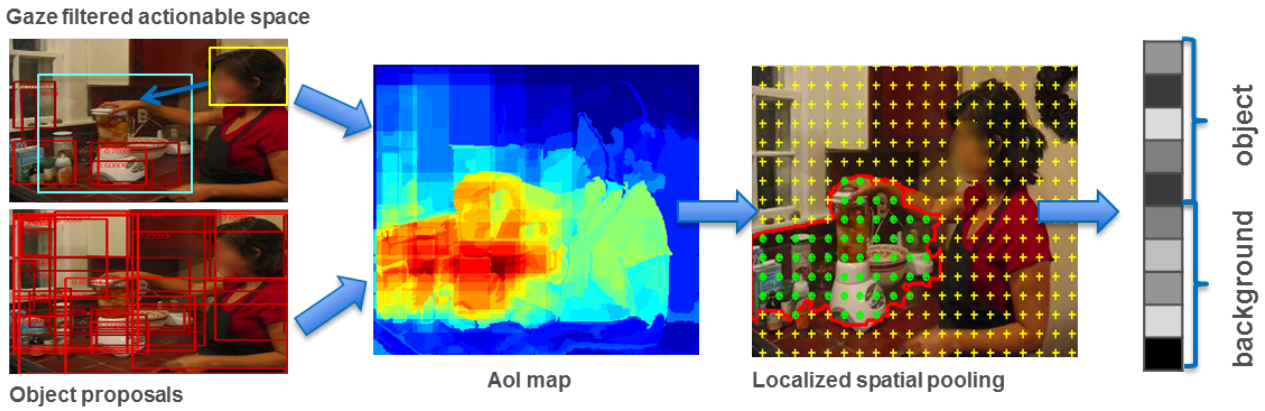

Figure 2: A walkthrough example on a cooking event in which the grasping hand and countertop objects are localized as Area of Interest. Best viewed in color.

Finally, objects pertaining to events are often ambiguous and are less likely to have rigid, well-defined boundaries.

Inspired by these developments, we propose an entity centric area of interest based feature pooling strategy[1]. We observe that some entities such as humans are easy to detect while event specific objects of interest are often hard to detect. Thus, by using the known entities as anchors, we first discover an Area of Interest (AoI) map in image keyframes and then use the AoI map to enhance event detection. Specifically, the AoI map is derived from the *visual salience* in the image which is weighted by the *actionable space* of the person involved in the event. We measure visual salience in terms of generic object-level region contrast and boundary completeness cues. The actionable space is defined as the space surrounding the human body that has potential for action [8], either by active interaction with the objects or by passive allocation of gaze-based attention. We model the actionable space and gaze direction using the detected human position and the face pose in the image. We posit that the visual and actionable spaces intersect to result in the AoI map that contains event-specific information. The AoI map is used to divide the image space into disparate regions from which features are pooled separately and concatenated into a single image signature. The main novelty of our paper is the visual-semantic feature pooling strategy that combines category independent, object-level salience in images with people-object interaction understanding in a single framework. We show that our approach is able to capture discriminative information that improves upon the state of the art for complex event detection.

2. OBJECTS IN ACTIONABLE SPACE

The pipeline of our approach are summarized in Figures 2 and 3. Given a video, we sample keyframes and densely extract patch-level features. Next, we sample a large number of rectangular regions based on object-level salience. We also perform face pose detection and use it to delineate the gaze and actionable space of the person in the event. The visual and the person-centric salience

[1]This work has been supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11-PC20066. The U.S. Government is au- thorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and con- clusions contained herein are those of the authors and should not be in- terpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/NBC, or the U.S. Government.

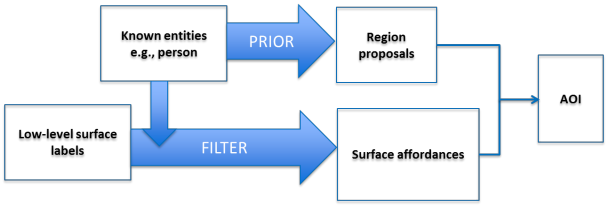

Figure 3: Overall pipeline

maps are combined into an AoI map, which is used for spatially localized feature pooling.

2.1 Object proposals

In our method, we bypass the notoriously difficult problem of precise object labeling and rely on a generic objectness measure to propose large number of object candidates in a bottom-up manner. We use the objectness algorithm [3], which quantifies how likely it is for an image window to contain an object of any class. The characteristics captured are general properties related to most objects, i.e., that objects appear different from their surroundings and having a closed boundary. They are modeled by a combination of multiple cues i.e., frequency domain multi-scale saliency, center-surround color contrast, boundary edge density and superpixel straddling (superpixels crossing over object windows), whose parameters have been optimized on the VoC PASCAL dataset. We found in our experiments that objectness adequately fires on many individual objects as well as on collections of objects of similar dimensions. This is useful for capturing the spatial extent of multiple objects of interest jointly, e.g., countertop items in Figure 2. In our pipeline, we sample a large number of windows from the objectness distribution according to non-maxima sampling procedure and rank them according to the scores. To generate an objectness map, we compute the pixel-wise objectness score by summing over all the windows that contains the pixel. Such a heat map is illustrated in Figure 2.

2.2 Human actionable space

The actionable space of a person involved in an event (i.e.,actor) is the area spatially surrounding the body and within the camera personś view. This is the area containing event-specific objects that the actor is most likely to influence. There is psychological evidence that objects within the actionable space are encoded in a body-centered reference frame [8]. Inspired by this theory, we de-

lineate the near space of the human into eight functional regions to form an actionable map, as shown in Figure 4. These regions encode objects based on their spatial relationship to the human body. The side regions R1 to R4 encode objects that appear in the space situated next to the actor, e.g., when "repairing an appliance". The below-torso regions R5 to R7 encode objects that appear when the actor is standing in front or behind a table top e.g., while "making a sandwich". The torso region R8 encodes objects that the actor holds close to the body e.g., an animal or a baby. The size of each region is proportional to estimated body dimensions. Based on the face size, we estimate the standing height and the arm span of the human, assuming standard adult body proportions [1]. Each region is delineated based on the extent of arm span and torso. Because of its dependence on body dimensions, the size of the regions depend on the camera's viewing distance; when the person is far from the camera, the regions cover a small area in the image and vice versa. Also, based on the human position and image dimensions, if the visibility of certain regions are below a threshold, these are turned off in the actionable map.

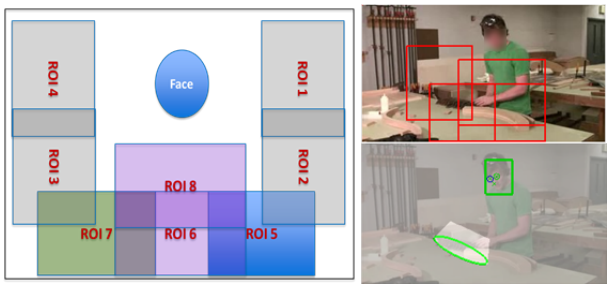

Figure 4: Left: Functional regions R1-R8 form the human actionable map. Right: Modulation by gaze direction.

2.3 Gaze based filtering

While performing an activity, one is likely to focus gaze at the objects of interest. To incorporate this bias spatially, we model the gaze direction of the actor using two types of distributions: (a) a *gaze map* that modulates the functional regions and (b) a *gaze volume* that modulates the objectness regions. The role of the gaze map is to selectively activate functional regions that intersect with the gaze direction. To model it, we consider the estimated Euler angles of face pose, i.e., the roll (up-down) and yaw (sideways) angles. We measure the angular deviations from frontal view (-90 to +90 degs.) and quantize each into five bins to generate a 5x5 grid. The responses over all keyframes are summed up and thresholded to find the quantized direction of maximum gaze in the video. Next, the role of the gaze volume is to selectively activate objectness regions. We model the gaze volume as a cone shaped distribution along the gaze ray, with the apex at the center of the eyes [10]. The gaze ray is a 3D ray emanating from the center of the eyes and perpendicular to the face plane. The cone shape of the gaze volume is simulated by concatenating 2D Gaussian distributions centered on gaze ray and of increasing covariances. This is then projected onto the 2D image space to find the gaze area of interest. The objectness regions that intersect the gaze area are selected for activation. The contour of a 1/3 gaze area is shown in Figure 4.

2.4 Multicue combination for AoI

In this step, we fuse the objectness with the actionable space, modulated by a gaze direction. Based on where the person looks, there are two cases to consider. (1) When the actor faces the camera, i.e., the center bin of the gaze map is maximum and the gaze area is a Gaussian that projects back on the person's face. Under these conditions, the gaze is non-informative, i.e., all the functional regions are equally likely. The objectness regions are activated if they intersect the actionable map (case 1 of Equation 1). (2) When the actor faces away from the camera. Based on which of the off-center columns of the gaze map are activated, the functional regions coincide with the gaze direction are selectively activated. For example, in Figure 4, regions R1,R2 and R5 are deactivated because the person looks leftwards. Under this condition, the gaze volume backprojects onto some region in the image. The objectness regions that intersect both this area and the actionable map are activated (case 2 of Equation 1). This combination rule for each pixel in the AoI map is mathematically summarized as follows.

$$G_m[3,3] = \begin{cases} 1 & \sum_{p:p\cap Fn} 1 \circ \sum_{p:p\cap Obj} o_i \\ \text{otherwise,} & \sum_{p:p\cap \hat{Fn}} 1 \circ \sum_{p:p\cap Obj} o_i \circ \sum_{p:p\cap Gaze} g_i \end{cases} \quad (1)$$

where the notations are $G_m[.]$ for gaze map coordinates, p is a pixel in the AoI map, Fn denotes all functional regions and $\hat{Fn}$ are the gaze coincident regions, oi and gi are objectness and gaze score, resp. The contours of AOI are illustrated for a few sample images in Figure 5.

3. EXPERIMENTS

3.1 Dataset

We evaluate our model on the person-object interaction clips derived from the NIST TRECVID corpus [2]. This dataset consists of video clips from seven events. The background dataset consists of 5000 miscellaneous videos clips sampled from non event videos. Each clip is about 9 secs. in duration. We process keyframes sampled at 2 frames per second. The number of videos from each category are as follows: background (2746), woodworking (138), blowing off candles (162), changing vehicle tire (68), grooming an animal (58), making a sandwich (82), repairing an appliance (73) and sewing (57). The dataset is evaluated over ten folds with 60% train and 40% test instances, generated by random sampling with replacement.

3.2 Data representation

We extract four types of features: (a) patch-level visual features extracted densely over the image grid, (b) objectness windows and (c) actionable map, (d) gaze map and volume. The processing time is about 2 secs. per keyframe. For patch-level representation, we consider two types of visual descriptors- Dense Trajectory Features with Histogram of Oriented Gradients (DTF-HoG) [15] and Dense Scale Invariant Feature Transform (Dense SIFT). We extract feature descriptors, quantize them using k-means to model a discrete codebook of visual words and finally encode the features into a fixed length vector per keyframe. We use a codebook vocabulary size of 10,000 words. A video feature is represented by the average over all the keyframe feature vectors.

For objectness, we use the publicly available software from the author's website [3]. We include the optional steps of color contrast and superpixel based processing along with multi-scale saliency within the sliding window mechanism. For the posterior object score, we set the Bernoulli probability of objectness $p(obj)$ to 0.10. We also use the author recommended 1000 windows for proper coverage. We tested non-maxima suppression (nms) and multinomial sampling of windows. We found that multinomial sampling tends to give better coverage, but extracts large windows while nms

sampling fired more frequently on medium sized objects. We chose nms because it was better suited for AoI localization in our case.

We apply Pittpatt software for face pose detection. The parameters are set to meet about 90% precision in face detection. In case of multiple face detections, we consider the center-most face for extracting the human centered actionable map. We assume human adult proportions with body height and arm length set to three and seven times the face size, resp. [1]. The gaze map is 5x5 spatial grid. For the gaze volume, we assume that the center of eyes is located at the origin and a point at a arbitrary distance on the gaze ray is at (0,0,z). The variance of gaze is a 3D Gaussian centered at $(0, 0, z) \sim N(0, \sigma_{x,y,z})$, which is backprojected to obtain a 2D distribution. We model the cone divergence angle as a function of the face size as well as the image size to incorporate the viewing distance in the estimation.

We perform one versus all max-margin classification to categorize each video into one of the seven events or background. We use libSVM [5] software with the histogram intersection kernel and a pre-defined penalty c as 1. The same settings were maintained across all the experiments for consistency. Unless otherwise mentioned, performance is measured by the mean average precision of the ranked list of SVM decision scores.

3.3 Comparison between SP-FP and AS-FP

We first compare image-centered spatial pyramid (SP-FP) and human-centered actionable space (AS-FP) for feature pooling. We consider a two-levelled pyramid. Level 1 is the standard visual word histogram from the whole image. Level 2 in the spatial pyramid is a symmetric 3x3 grid. In the actionable space, level 2 consists of the functional region partitions R1 to R8 (Fig. 3). If a particular region is absent, its contribution is set to zero. We pool dense SIFT features and normalize them to generate independent feature histograms per region. Next, we concatenate them with level 0 histogram and compute the video average. These video features are inputs to the classifier. The results are shown in Table 1. Overall, our average accuracy of 44.4% is about 3% higher than SP-FP. However, there is varying influence of the pooling strategy across events. "Woodworking", "blowing off candles" (renamed birthday, for short) and "making sandwich" benefit most from human-centered pooling, with a gain of about 7%. This could be attributed to a distinct relative spatial geometry that exists between person and objects relevant to these events which can not be captured by the symmetric image-centered geometry of spatial pyramid.

events		woodwork	birthday	tire	animal
Dense SIFT	SP-FP	63	60	14	7
	AS-FP	**66**	**65**	14	**9**
events		**sandwich**	**appliance**		**sewing**
Dense SIFT	SP-FP	52	45		55
	AS-FP	**57**	**46**		54

Table 1: Average precision scores comparing proposed actionable space (Figure 4) and spatial pyramid [9] based feature pooling.

3.4 Comparison between AOI, Dense and Random sampling techniques

It is possible that the above approach over-partitions the feature space, which leads to sub-optimal performance. To verify this intuition, in the next experiment we combine gaze and objectness cues with the actionable map and partition the feature space into two regions, an AoI foreground and a background region. The features within each region are pooled separately and concatenated. To evaluate the performance of this bi-region pooling, we also compare against two baselines: dense pooling (Dense) and random spatial pooling (Rand.) [12]. Dense representation is the typical average pooling of bag of features for the whole video. For random representation, we pool 10K randomly sampled patches per video. We also compare the efficacies of two different features: DTF-HoG and Dense SIFT. The results are shown in Tables 2 and 3.

Figure 5: Contours of AOI map modulated by object proposals, gaze and actionable space based filtering.

events		woodwork	birthday	tire	animal
DTF-HOG	Dense	27±4	25±2	4±1	7±1
	AOI	**28**±5	**47**±5	**8**±1	**8**±2
events		**sandwich**	**appliance**		**sewing**
DTF-HOG	Dense	12±3	22±3		9±1
	AOI	**25**±5	**22**±2		**9**±2

Table 2: Comparison between AOI and dense pooling using DTF-HoG

Overall, the performance of Dense SIFT, which represents texture and shape, exceeds that of DTF-HoG, which captures motion. This may be attributed to the inability of motion features to capture subtle movements in person-object interactions. Also, the overall standard deviation across 10 folds is low. We observe that "woodworking" and "making sandwich" always benefit from spatially localized pooling. For other events, there is great increase in accuracy when AOI based pooling is applied. The largest AP gain, greater than 100%, is seen for "grooming animal" which increases from 9% to 20%, followed by "repairing appliance" (54%), "changing tire" (23%) and "blowing off candles" (13%)[2]. This could be attributed

[2] $AP_{gain} = (AP_{final}\text{-}AP_{initial})/AP_{initial}$

events		woodwork	birthday	tire	animal
Dense SIFT	Dense	43±3	61±5	13±2	9±1
	AOI	**53**±5	**69**±2	**16**±3	**20**±2
	Rand.	41±2	49±3	12±1	10±2

events		sandwich	Appliance	sewing
Dense SIFT	Dense	46±4	35±2	61±2
	AOI	**50**±3	**54**±3	**59**±3
	Rand.	31±1	43±2	48±4

Table 3: Comparison between AOI, dense and random pooling using Dense SIFT.

to the focussed search performed by our method for relevant objects in the vicinity of the person. "Sewing" is the only event that shows negative performance under spatial pooling. This could be due to a large intra-class variation, where a person is involved in either using a sewing machine, cloth cutting or displaying garments.

events		woodwork	birthday	tire	animal	sandwich	appliance	sewing
ρ(gaze, precision)	AOI	**-0.28**	**0.28**	**0.34**	**0.30**	**0.18**	**0.24**	**0.21**
	Dense	-0.07	0.08	-0.01	-0.02	0.06	0.03	-0.09

Table 4: Pearson correlation between informative gaze and precision @ top 50 detections. This implies that for AoI pooling, the highly ranked correctly detected videos are more likely to contain informative gaze.

events		woodwork	birthday	tire	animal	sandwich	appliance	sewing
ρ(w_fore, w_bkgnd)	AOI	**0.07**	**0.06**	**0.03**	**0.01**	**0.01**	**0.00**	**0.25**

Table 5: Pearson correlation between learned classifier weights of corresponding dimensions from AoI and background. There is negligible correlation between weights of the same visual words.

3.5 Impact of AoI priors

We provide a novel way of analyzing the contributions of different priors used in our model. Specifically, we study the classifier outputs vis-a-vis the gaze and AoI prior assumptions. Table 4 shows the measure of Pearson correlation between informative gaze prior and precision at top 50 detections. A gaze pattern is informative if it is used in the AoI region selection (Sec. 2.3). We observe that there is a direct correlation (mostly positive) between availability of informative gaze and correct detections in the AoI based approach, whilst the same prior is uncorrelated with the dense detections. We can conclude that gaze drives proper selection of AoI, which in turn improves event detection. Next, Table 5 shows the correlation between the learnt classifier weights for corresponding features appearing in the AoI and background dimensions. I.e., since the AoI pooling is formed by concatenation, the same features are considered twice, once as part of the AoI and second, as part of the background. Hence, the correlation measure should verify that the same features are not simply reinforced across the two dimensions. We see that there is negligible correlation between weights of the same visual words, from which we may conclude that complementary information is learnt from the two spaces, leading to an improvement in event detection.

4. CONCLUSION

In this paper, we proposed a human centric region of interest detection and visual-semantic pooling scheme for complex event detection. We discovered an Area of Interest (AoI) map in image keyframes, which is used for differential pooling of features. Our experiments shows that our proposed semantic feature pooling is able to surpass the performance of state of the art approaches. We also show that the method's high accuracy is directly correlated with the human actionable priors which are used for differential pooling. In the future, we hope to extend our approach to include a wider range of entities to spatial ground the search for discriminative information in event detection.

5. REFERENCES

[1] Wikipedia entry on body proportions.

[2] Trecvid multimedia event detection track, 2011.

[3] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *PAMI*, 34(11):2189–2202, November 2012.

[4] T. Althoff, H. O. Song, and T. Darrell. Detection bank: an object detection based video representation for multimedia event recognition. In *ACM Multimedia*, pages 1065–1068. ACM, 2012.

[5] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[6] A. Fathi, A. Farhadi, and J. Rehg. Understanding egocentric activities. In *ICCV*, pages 407–414, 2011.

[7] Y. Jiang, J. Yuan, and G. Yu. Randomized spatial partition for scene recognition. In *ECCV*, pages II: 730–743, 2012.

[8] L. Lamport. *The Oxford Handbook of Thinking and Reasoning: Visuospatial Thinking*. Oxford University Press, 2012.

[9] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, pages II: 2169–2178, 2006.

[10] H. S. Park, E. Jain, and Y. Sheikh. 3d social saliency from head-mounted cameras. In *NIPS*, December 2012.

[11] J. Sanchez, F. Perronnin, and T. de Campos. Modeling the spatial layout of images beyond spatial pyramids. *PRL*, 33(16):2216–2223, December 2012.

[12] F. Shi, E. Petriu, and R. Laganiere. Sampling strategies for real-time action recognition. In *CVPR*, pages 2595–2602, 2013.

[13] A. Tamrakar, S. Ali, Q. Yu, J. Liu, O. Javed, A. Divakaran, H. Cheng, and H. Sawhney. Evaluation of low-level features and their combinations for complex event detection in open source videos. In *CVPR*, pages 3681–3688, 2012.

[14] J. Uijlings, A. Smeulders, and R. Scha. What is the spatial extent of an object? In *CVPR*, pages 770–777, 2009.

[15] H. Wang, A. Klaser, C. Schmid, and C. Liu. Action recognition by dense trajectories. In *CVPR*, pages 3169–3176, 2011.