A survey on computer vision tools for action recognition, crowd surveillance and suspect retrieval

Teófilo E. de Campos[1]

[1]Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, GU2 7XH, UK

t.decampos@st-annes.oxon.org
http://personal.ee.surrey.ac.uk/Personal/T.Decampos/

Abstract. *This paper briefly surveys computer vision tools that can be used on surveillance videos for crowd monitoring in order to detect anomalous events and retrieve suspects. Focus is given to methods for action recognition and event detection.*

(a) Satellite image, indicating the location of the event. ©Igonográfico, Globo.

(b) Still image from the surveillance camera that captured a flying toilet. ©Diário de Pernanbuco.

(c) Crime scene with a shattered toilet seat. ©Carlos Ezequiel Vannoni/Agência JCM/Fotoarena, Folha de São Paulo.

(d) Suspect arrested in less than 72hs with support from the community, thanks to a reward offered for helpful leads. ©Lucas Liausu, Globo.

Figure 1. In May 2014, hooligans dropped a toilet from the top deck of a stadium onto the street below, where it struck and killed a supporter of the rival team in Recife, Brazil.

1. Introduction

Monitoring large crowds is a very challenging task, which currently is done using surveillance cameras controlled manually by remote human operators. The number of video

feeds is usually overwhelmingly large for the number of officers who monitor them, rendering such surveillance systems almost useless for real-time detection of threats. To add to this challenge, the most relevant events are anomalous, i.e., they are rare and last for a very short period. It is often hard to obtain realistic labelled data for such events and it can be impossible to specify how they happen. Criminal activities usually do not follow a well known pattern. In the example of Figure 1, hooligans threw toilet seats out of the top deck of a football stadium. One cannot expect vision researchers to design methods to detect gangs carrying toilet seats, but some actions, events and crowd patterns could trigger alarms.

This paper presents a brief overview of computer vision techniques that can be used help monitoring surveillance videos. Also included are methods to help retrieving suspects in smart cities and mining social networks to help retrieving further information.

From Sections 2 to 4, this paper reviews methods with decreasing order of specificity w.r.t. how the threat is modelled, i.e., with increasing levels of generalisation for practical applications. Next, Section 5 reviews a tool that helps with the investigation of crimes in crowded environment and smart cities. A discussion is presented in Section 6.

2. Detecting events and actions in video

Video-based activity recognition and event detection is a very widely explored research field in computer vision [Poppe 2010, Ke et al. 2013], though it remains a very challenging problem, with a wide range of applications. In the remaining of this section I briefly highlight some of the works in this area.

2.1. The focused approach: human-centred action detection

One possible approach to crowd analysis is to use human detectors and try to describe the action of the set of most salient people in the video sequences. For instance, a part-based method such as [Ikisler and Forsyth 2007, Ramanan et al. 2007, Shotton et al. 2011] can be used to locate each body part. Once the parts are found, their trajectories can be processed using methods like HMM to describe actions. The problem of those approaches is that they tend not to be reliable in crowded scenes. Perhaps a more robust technique is to use bounding boxes around people (that can be found using methods such as those of [Dalal and Triggs 2005, Felzenswalb et al. 2009]) and describing the motion within those boxes in a discriminative way, without necessarily locating each limb of the person. For example, [Gorelick et al. 2007] rely on segmentation of the actor to build a 3D (space-time) shape that can be described as a vector and matched to templates. [de Campos et al. 2011] and [Kläser et al. 2010] also describe actions as space-time volumes, but build feature vectors based on a 3D generalisation of HoG [Kläser et al. 2008], so they do require actor detection and tracking, but segmentation is not needed, enabling these methods to be applied in more realistic videos. [Ke et al. 2009] proposed to over segment images in a video sequence, obtaining volumetric parts which are augmented using flow. Rather than treating an event template as an atomic entity, they separately match by parts in space and time. This yields further robustness to clutter in crowded scenes.

2.2. The “blind” approach: pooling of features and action banks

In cluttered video, it is harder to rely on human detectors. A more robust approach is to use methods that describe video sequences as an unordered set of local

space-time features, i.e., as bags of visual words (BoW) or, more generically, pooling methods. [Wang et al. 2009] presented a benchmark of local space-time features for the BoW framework, concluding that densely sampled features gave the best results. The best local feature extraction methods were HOG3D [Kläser et al. 2008] and HOG/HOF [Laptev et al. 2008]. Better results were obtained more recently using dense trajectories [Wang et al. 2013].

[Chatfield et al. 2011] presented a benchmark on coding methods for methods for feature pooling on static images, concluding that Fisher Kernels [Sánchez et al. 2013] are the best. The use of Fisher Kernels coding for action classification was demonstrated on two small datasets in [Atmosukarto et al. 2012]. [Oneata et al. 2014] presented state-of-the-art results on action classification and temporal localisation on challenging datasets using Fisher vectors.

In [Wang et al. 2013], the authors combined dense trajectories and motion boundaries to build action descriptors, i.e., a hybrid between the methods of this section and those of previous section.

A more promising alternative is to represent video sequences as action banks, i.e., instead of pooling local space-time features, combine local event detectors to build global feature vectors. This approach is computationally more demanding than the methods above, but gave much better results in challenging benchmarks [Sadanand and Corso 2012].

In [Chatfield et al. 2014] the authors showed that Convolutional Neural Networks lead to significant better results than other coding methods on static image datasets. To the best of my knowledge, this approach has not yet been evaluated for action classification or event detection in video sequences.

3. Tracking-based methods

Given tracking results, i.e., a sequence of coordinates, higher level algorithms can be used in order to detect and classify events using methods such as HMMs [Almajai et al. 2010], or structured learning [Yan et al. 2012]. This has a wide range of applications, from shop layout optimisation to security.

3.1. Detection of abandoned objects

Following demand from the police, the detection of some specific actions in surveillance video has become a hot topic, with abandoned bags detection topping the list. Much of the work is based on heuristics that are applied to the results of background subtraction and segmentation. This was used in [Tian et al. 2008] to detect static regions and classify them as abandoned or removed objects. Such techniques are simpler than those used for action classification, but can effectively be applied in real-time with static surveillance cameras. However, they may trigger too many false alarms.

In [Fan et al. 2013], the authors proposed to reduce the number of false positives by representing abandoned object alerts by relative attributes, namely staticness, foregroundness and abandonment (Figure 2). The relative strengths of these attributes are quantified using a ranking function learnt on low-level spatial and temporal features. With these features, they apply a linear ranking algorithm to sort alerts according to their relevance to the end-user.

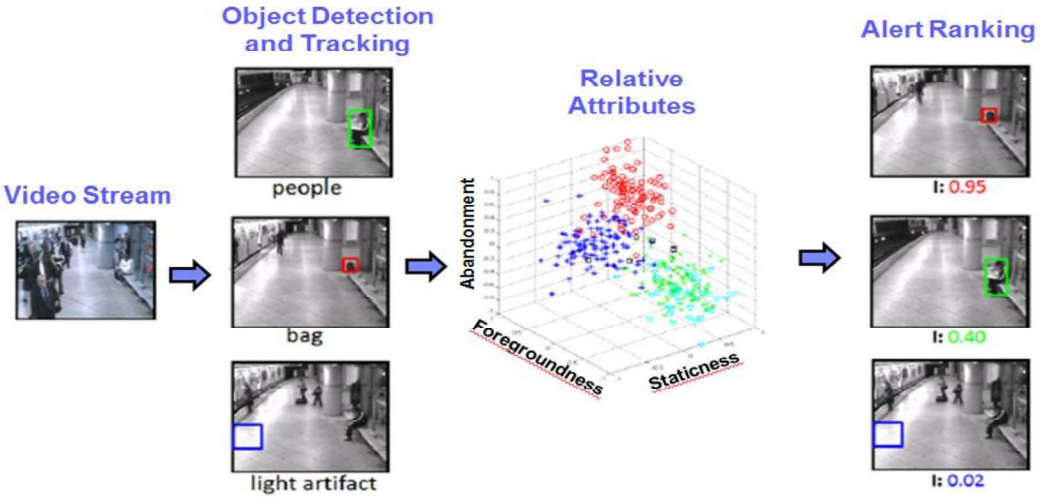

Figure 2. Method for detection of abandoned objects outlined by [Fan et al. 2013] ©Rogerio Feris.

3.2. Crowd analysis

In many camera setups, it is not possible to monitor crowds based on the detection of actions of individuals, because there might be too much occlusion and the number or pixels that a person occupies in the image may be too small. A straightforward extension of techniques designed for non-crowded scenes cannot be suitable for dealing with crowded situations [Jacques-Jr et al. 2010]. A survey of this field was presented in [Thida et al. 2013]. The rest of this section reviews some recent methods.

[Rodriguez et al. 2011] "adhere to the insight that despite the fact that the entire space of possible crowd behaviours is infinite, the space of distinguishable crowd motion patterns may not be all that large."[1] Their method learns a set of crowd behaviour priors from videos gathered from the Internet. During testing, crowd patches are matched to the database in a similar fashion to that done for data-driven image denoising and inpainting (see Figure 3). Therefore, this method requires extensive searching of similar patches in the database, while making a strong assumption that the motion of individuals in a particular query patch can be found in the database.

[Idrees et al. 2014] rely completely on information that is readily available in the sequences. They use automatic identification of prominent individuals from the crowd that are easy to track and model the behaviour of individuals in a dense crowd using Neighborhood Motion Concurrence. This predicts the position of an individual based on the motion of its neighbours. When the individual moves with the crowd flow, this method predicts motion while leveraging five-frame instantaneous flow in case of dynamically changing flow and anomalies.

Ground truthing crowd analysis algorithms to train and evaluate algorithms is particularly challenging, as one might need to annotate the trajectory of every individual in the scene. A possible solution is to resort to datasets synthesised by crowd simulation algorithms, such as the Agoraset [Courty et al. 2014], thought they may not be perfectly

[1]See `http://www.di.ens.fr/willow/research/datadriven/` for demonstrations.

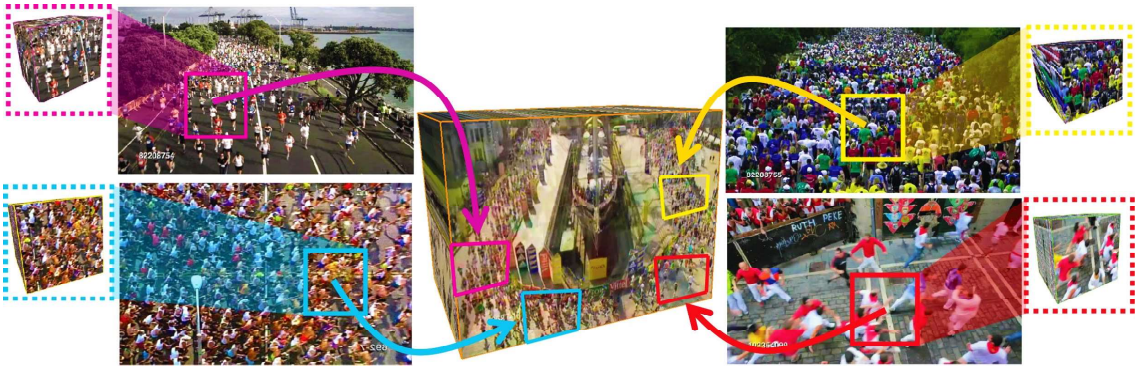

Figure 3. [Rodriguez et al. 2011] proposed to depict crowded scenes (like the one in the centre) as a combination of previously observed crowd patches. Each crowd patch contains a particular combination of crowd behaviour patterns (people running in a particular direction in this example) ©IEEE.

realistic, not only in terms of the visual appearance of the crowd, but also because they may not appropriately model abnormal behaviours.

4. Detecting unknown patterns with anomaly detection methods

As outlined in the previous sections, the research literature is vast in terms of specific and isolated problems, such as detecting a pre-defined set of gestures, actions and events. When demand is high enough, vision researchers tend to focus on specific problems and propose bespoke solutions that can work in real-time, such as those for abandoned object detection or detection of checkout frauds and sweethearting in shops [Fan et al. 2009]. However, little has been done on detection of unexpected events, such as that of Figure 1. Perhaps the method proposed in [Endres et al. 2011] could be applied to that scenario, where hooliganism (e.g. brawl in stadiums) is detected using saliency of optical flow patterns. Early warnings of anomalous events can minimise the damage caused by criminal activities.

A common approach for anomaly detection is to combine information from the motion of individuals with that of the motion of groups. [Leach et al. 2014] implement an unsupervised context-aware process that utilises both scene and social contexts, i.e., priors are built from typical trajectories in a scene and social groups are detected to give further context for anomalous behaviour detection (e.g. loitering). [Cho and Kang 2014] model individual and group behaviour using a hybrid agent system that includes static and dynamic agents to observe efficiently the corresponding individual and interactive behaviours in a crowded scene. The behaviour of a crowd is modelled as a bag of words through the integration of static and dynamic agent information to determine abnormalities.

In [Chong et al. 2014], the authors proposed to use hierarchical Dirichlet processes to model the motion of regions of interest at global and local levels. Anomalies on both levels are detected as events for which statistical features and location are beyond the normal expected range according to the learnt templates.

The above methods were designed following assumptions about people's motion and their interactions in video sequences. However, they were tested in datasets where a pre-defined set of anomalous events occur in the test set. It is hard to tell if they would

generalise well to unexpected events. In [Kittler et al. 2014], the authors present a survey of the anomaly detection literature and propose a framework for contextual (domain) anomaly detection in video sequences. However, that framework is still to be evaluated in generic crowd surveillance videos.

5. Searching for suspects using semantic attributes

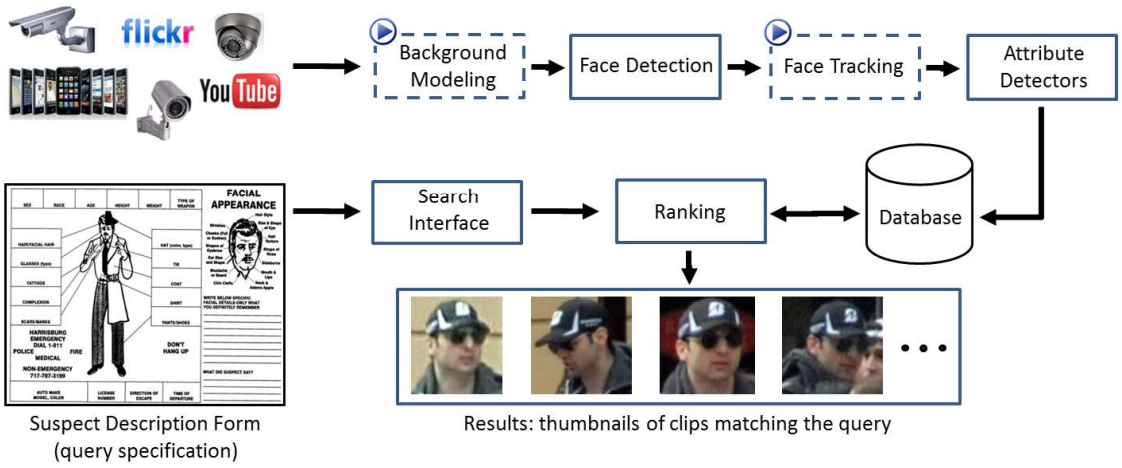

Figure 4. Attribute-based retrieval system of [Feris et al. 2014] ©ACM.

In practical applications where surveillance cameras help with crime investigations, queries are made by semantic attributes, e.g. "show me all people with a beard and sunglasses, wearing a white hat and a patterned blue shirt, from all metro cameras in the downtown area, from 2pm to 4pm last Saturday" [Feris et al. 2014]. The IBM system (Figure 4), currently deployed by many police departments around the world, combines not only surveillance cameras, but also images and videos from Flickr, YouTube and images captured from mobile phones. It uses adaptive background subtraction (for static cameras), face detection, face tracking and attribute detectors. The latter combines a battery of detectors (bald, beard, clothing colour, etc) on faces, torso and legs regions. Queries are textual and the results are ranked according to the relevance of the multi-attribute query, using the "learning to rank" approach [Siddiquie et al. 2011].

6. Discussion

This paper briefly surveyed computer vision tools that can be applied to increase the security of large scale events, such as football matches. Focus was given to methods for action recognition, event detection and crowd surveillance. I also summarised a state-of-the-art method for attribute-based person retrieval from large scale surveillance systems.

Although the research community in these areas is large, much of the work is focused on obtaining good results on public benchmarks, rather than actually solving real problems, where it can be impossible to specify a model of anomalous actions and events. However, anomaly detection research is emerging and shall have direct impact on the society in the near future.

In addition, data gathered from surveillance cameras will prove a lot more effective in crime prevention and crime investigation when vision systems start to integrate

data captured live with data obtained from the web and from social networks. Vision researchers have already been focusing their efforts on face recognition and image labelling on social networks [Sukhbaatar and Fergus 2014, McAuley and Leskovec 2012]. Furthermore, visual information from social network must be integrated with other social network data that enables things like preference estimation [McAuley and Leskovec 2013], personality categorisation [Kosinski et al. 2014], etc.

Acknowledgements

I would like to thank the International Relations office at FEPS/Surrey and the University Global Partnership Network (UGPN) for enabling an academic visit to Maria da Graça Campos Pimentel, as that visit triggered discussions that lead to this paper. My work is currently sponsored by the S3A project, grant EP/L000539/1 from the EPSRC, United Kingdom and by the ROMEO project, FP7 grant 287896 from the European Commission. I am also grateful for the SEMISH organisers, for sponsoring my attendance to the SBC conference.

References

[Almajai et al. 2010] Almajai, I., Kittler, J., DeCampos, T., Christmas, W., Yan, F., Windridge, D., and Khan, A. (2010). Ball event recognition using hmm for automatic tennis annotation. In *Proceedings of Intl. Conf. on Image Processing (ICIP)*.

[Atmosukarto et al. 2012] Atmosukarto, I., Ghanem, B., and Ahuja, N. (2012). Trajectory-based fisher kernel representation for action recognition in videos. In *21st International Conference on Pattern Recognition (ICPR)*, pages 3333–3336.

[Chatfield et al. 2011] Chatfield, K., Lempitsky, V., Vedaldi, A., and Zisserman, A. (2011). The devil is in the details: an evaluation of recent feature encoding methods. In *British Machine Vision Conference*.

[Chatfield et al. 2014] Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. Technical report, University of Oxford. Archived in arXiv 1405.3531.

[Cho and Kang 2014] Cho, S.-H. and Kang, H.-B. (2014). Abnormal behavior detection using hybrid agents in crowded scenes. *Pattern Recognition Letters*, 44:64–70.

[Chong et al. 2014] Chong, X., Liu, W., Huang, P., and Badler, N. I. (2014). Hierarchical crowd analysis and anomaly detection. *Journal of Visual Languages & Computing*. `http://dx.doi.org/10.1016/j.jvlc.2013.12.002i`.

[Courty et al. 2014] Courty, N., Allain, P., Creusot, C., and Corpetti, T. (2014). Using the agoraset dataset: assessing for the quality of crowd video analysis methods. *Pattern Recognition Letters*.

[Dalal and Triggs 2005] Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proc IEEE Conf on Computer Vision and Pattern Recognition, San Diego CA, June 20-25*.

[de Campos et al. 2011] de Campos, T., Barnard, M., Mikolajczyk, K., Kittler, J., Yan, F., Christmas, W., and Windridge, D. (2011). An evaluation of bags-of-words and spatio-temporal shapes for action recognition. In *IEEE Workshop on Applications of Computer Vision (WACV)*, Kona, Hawaii.

[Endres et al. 2011] Endres, D., Neumann, H., Kolesnik, M., and Giese, M. (2011). Hooligan detection: the effects of saliency and expert knowledge. In *4th International Conference on Imaging for Crime Detection and Prevention (ICDP)*, pages 1–6. IET.

[Fan et al. 2009] Fan, Q., Bobbitt, R., Zhai, Y., Yanagawa, A., Pankanti, S., and Hampapur, A. (2009). Recognition of repetitive sequential human activity. In *Proc of the IEEE Conf on Computer Vision and Pattern Recognition*, pages 943–950.

[Fan et al. 2013] Fan, Q., Gabbur, P., and Pankanti, S. (2013). Relative attributes for large-scale abandoned object detection. In *Proc 14th Int Conf on Computer Vision, Australia*, pages 2736–2743.

[Felzenswalb et al. 2009] Felzenswalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2009). Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Feris et al. 2014] Feris, R., Bobbitt, R., Brown, L., and Pankanti, S. (2014). Attribute-based people search: Lessons learnt from a practical surveillance system. In *Proceedings of International Conference on Multimedia Retrieval (ICMR)*. ACM.

[Gorelick et al. 2007] Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. (2007). Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253.

[Idrees et al. 2014] Idrees, H., Warner, N., and Shah, M. (2014). Tracking in dense crowds using prominence and neighborhood motion concurrence. *Image and Vision Computing*, 32(1):14–26.

[Ikisler and Forsyth 2007] Ikisler, N. and Forsyth, D. (2007). Searching video for complex activities with finite state models. In *Proc of the IEEE Conf on Computer Vision and Pattern Recognition*.

[Jacques-Jr et al. 2010] Jacques-Jr, J. C. S., Musse, S. R., and Jung, C. R. (2010). Crowd analysis using computer vision techniques. *IEEE Signal Processing Magazine*, 27(5):66–77.

[Ke et al. 2013] Ke, S.-R., Thuc, H. L. U., Lee, Y.-J., Hwang, J.-N., Yoo, J.-H., and Choi, K.-H. (2013). A review on video-based human activity recognition. *Computers*, 2(2):88–131.

[Ke et al. 2009] Ke, Y., Sukthankar, R., and Herbert, M. (2009). Event detection in crowded videos. In *Proc 12th Int Conf on Computer Vision, Kyoto, Japan, Sept 27 - Oct 4*.

[Kittler et al. 2014] Kittler, J., Christmas, W., de Campos, T., Windridge, D., Yan, F., Illingworth, J., and Osman, M. (2014). Domain anomaly detection in machine perception: A system architecture and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5):845–859. `http://dx.doi.org/10.1109/TPAMI.2013.209`.

[Kläser et al. 2008] Kläser, A., Marszałek, M., and Schmid, C. (2008). A spatio-temporal descriptor based on 3D-gradients. In *British Machine Vision Conference*, pages 995–1004.

[Kläser et al. 2010] Kläser, A., Marszałek, M., Schmid, C., and Zisserman, A. (2010). Human focused action localization in video. In *International Workshop on Sign, Gesture, Activity*. (best paper award winner) in conjunction with ECCV.

[Kosinski et al. 2014] Kosinski, M., Bachrach, Y., Kohli, P., Stillwell, D., and Graepel, T. (2014). Manifestations of user personality in website choice and behaviour on online social networks. *Machine Learning*, 95(3):357–380.

[Laptev et al. 2008] Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In *Proc of the IEEE Conf on Computer Vision and Pattern Recognition*, pages 1–8.

[Leach et al. 2014] Leach, M. J., Sparks, E., and Robertson, N. M. (2014). Contextual anomaly detection in crowded surveillance scenes. *Pattern Recognition Letters*, 44(0):71–79. Pattern Recognition and Crowd Analysis.

[McAuley and Leskovec 2012] McAuley, J. J. and Leskovec, J. (2012). Image labeling on a network: using social-network metadata for image classification. In *Proc European Conf on Computer Vision*.

[McAuley and Leskovec 2013] McAuley, J. J. and Leskovec, J. (2013). From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *World Wide Web*, pages 897–908.

[Oneata et al. 2014] Oneata, D., Verbeek, J., and Schmid, C. (2014). Efficient Action Localization with Approximately Normalized Fisher Vectors. In *Proc of the IEEE Conf on Computer Vision and Pattern Recognition*, Columbus, OH, United States.

[Poppe 2010] Poppe, R. (2010). A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990.

[Ramanan et al. 2007] Ramanan, D., Forsyth, D. A., and Zisserman, A. (2007). Tracking people by learning their appearance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):65–81.

[Rodriguez et al. 2011] Rodriguez, M., Sivič, J., Laptev, I., and Audibert, J.-Y. (2011). Data-driven crowd analysis in videos. In *Proc 13th Int Conf on Computer Vision, Barcelona, Spain*.

[Sadanand and Corso 2012] Sadanand, S. and Corso, J. J. (2012). Action bank: A high-level representation of activity in video. In *Proc of the IEEE Conf on Computer Vision and Pattern Recognition*, pages 1234–1241.

[Sánchez et al. 2013] Sánchez, J., Perronnin, F., Mensink, T., and Verbeek, J. (2013). Image classification with the fisher vector: Theory and practice. *International journal of computer vision*, 105(3):222–245.

[Shotton et al. 2011] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. (2011). Real-time human pose recognition in parts from single depth images. In *Proc of the IEEE Conf on Computer Vision and Pattern Recognition*.

[Siddiquie et al. 2011] Siddiquie, B., Feris, R. S., and Davis, L. S. (2011). Image ranking and retrieval based on multi-attribute queries. In *Proc of the IEEE Conf on Computer Vision and Pattern Recognition*, pages 801–808.

[Sukhbaatar and Fergus 2014] Sukhbaatar, S. and Fergus, R. (2014). Learning from noisy labels with deep neural networks. *arXiv preprint arXiv:1406.2080*.

[Thida et al. 2013] Thida, M., Yong, Y. L., Climent-Pérez, P., Eng, H.-l., and Remagnino, P. (2013). A literature review on video analytics of crowded scenes. In *Intelligent Multimedia Surveillance*, pages 17–36. Springer.

[Tian et al. 2008] Tian, Y. L., Feris, R. S., and Hampapur, A. (2008). Real-time detection of abandoned and removed objects in complex environments. In *VS*.

[Wang et al. 2013] Wang, H., Kläser, A., Schmid, C., and Liu, C.-L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1):60–79.

[Wang et al. 2009] Wang, H., Ullah, M. M., Kläser, A., Laptev, I., and Schmid, C. (2009). Evaluation of local spatio-temporal features for action recognition. In *Proc 20th British Machine Vision Conf, London, Sept 7-10*.

[Yan et al. 2012] Yan, F., Kittler, J., Mikolajczyk, K., and Windridge, D. (2012). Automatic annotation of court games with structured output learning. In *21st International Conference on Pattern Recognition (ICPR)*, pages 3577–3580. IEEE.