frontiers in Robotics and AI

PERSPECTIVE
published: 19 November 2015
doi: 10.3389/frobt.2015.00029

Object Detection: Current and Future Directions

Rodrigo Verschae[1*†] *and Javier Ruiz-del-Solar*[1,2]

[1] *Advanced Mining Technology Center, Universidad de Chile, Santiago, Chile,* [2] *Department of Electrical Engineering, Universidad de Chile, Santiago, Chile*

Object detection is a key ability required by most computer and robot vision systems. The latest research on this area has been making great progress in many directions. In the current manuscript, we give an overview of past research on object detection, outline the current main research directions, and discuss open problems and possible future directions.

Keywords: object detection, perspective, mini review, current directions, open problems

OPEN ACCESS

Edited by:
Venkatesh Babu Radhakrishnan, Indian Institute of Science Bangalore, India

Reviewed by:
Juxi Leitner, Queensland University of Technology, Australia
George Azzopardi, University of Groningen, Netherlands
Soma Biswas, Indian Institute of Science Bangalore, India

****Correspondence:***
Rodrigo Verschae
rodrigo@verschae.org

†Present address:
Rodrigo Verschae, Graduate School of Informatics, Kyoto University, Kyoto, Japan

Specialty section:
This article was submitted to Vision Systems Theory, Tools and Applications, a section of the journal Frontiers in Robotics and AI

Received: *20 July 2015*
Accepted: *04 November 2015*
Published: *19 November 2015*

Citation:
Verschae R and Ruiz-del-Solar J (2015) Object Detection: Current and Future Directions. Front. Robot. AI 2:29. doi: 10.3389/frobt.2015.00029

1. INTRODUCTION

During the last years, there has been a rapid and successful expansion on computer vision research. Parts of this success have come from adopting and adapting machine learning methods, while others from the development of new representations and models for specific computer vision problems or from the development of efficient solutions. One area that has attained great progress is object detection. The present works gives a *perspective on object detection research*.

Given a set of object classes, *object detection* consists in *determining the location and scale of all object instances, if any, that are present in an image*. Thus, the objective of an object detector is to find all object instances of one or more given object classes regardless of scale, location, pose, view with respect to the camera, partial occlusions, and illumination conditions.

In many computer vision systems, object detection is the first task being performed as it allows to obtain further information regarding the detected object and about the scene. Once an object instance has been detected (e.g., a face), it is be possible to obtain further information, including: (i) to recognize the specific instance (e.g., to identify the subject's face), (ii) to track the object over an image sequence (e.g., to track the face in a video), and (iii) to extract further information about the object (e.g., to determine the subject's gender), while it is also possible to (a) infer the presence or location of other objects in the scene (e.g., a hand may be near a face and at a similar scale) and (b) to better estimate further information about the scene (e.g., the type of scene, indoor versus outdoor, etc.), among other contextual information.

Object detection has been used in many applications, with the most popular ones being: (i) human-computer interaction (HCI), (ii) robotics (e.g., service robots), (iii) consumer electronics (e.g., smart-phones), (iv) security (e.g., recognition, tracking), (v) retrieval (e.g., search engines, photo management), and (vi) transportation (e.g., autonomous and assisted driving). Each of these applications has different requirements, including: processing time (off-line, on-line, or real-time), robustness to occlusions, invariance to rotations (e.g., in-plane rotations), and detection under pose changes. While many applications consider the detection of a single object class (e.g., faces) and from a single view (e.g., frontal faces), others require the detection of multiple object classes (humans, vehicles, etc.), or of a single class from multiple views (e.g., side and frontal view of vehicles). In general, most systems can detect only a single object class from a restricted set of views and poses.

Several surveys on detection and recognition have been published during the last years [see Hjelmås and Low (2001), Yang et al. (2002), Sun et al. (2006), Li and Allinson (2008), Enzweiler and Gavrila (2009), Dollar et al. (2012), Andreopoulos and Tsotsos (2013), Li et al. (2015), and Zafeiriou et al. (2015)], and there are four main problems related to object detection. The first one is *object localization*, which consists of determining the location and scale of a single object instance known to be present in the image; the second one is *object presence classification*, which corresponds to determining whether at least one object of a given class is present in an image (without giving any information about the location, scale, or the number of objects), while the third problem is *object recognition*, which consist in determining if a specific object instance is present in the image. The fourth related problem is *view and pose estimation*, which consist of determining the view of the object and the pose of the object.

The problem of *object presence classification* can be solved using object detection techniques, but in general, other methods are used, as determining the location and scale of the objects is not required, and determining only the presence can be done more efficiently. In some cases, *object recognition* can be solved using methods that do not require detecting the object in advance [e.g., using methods based on Local Interest Points such as Tuytelaars and Mikolajczyk (2008) and Ramanan and Niranjan (2012)]. Nevertheless, solving the object detection problem would solve (or help simplifying) these related problems. An additional, recently addressed problem corresponds to *determining the "objectness"* of an image patch, i.e., measuring the likeliness for an image window to contain an object of any class [e.g., Alexe et al. (2010), Endres and Hoiem (2010), and Huval et al. (2013)].

In the following, we give a summary of past research on object detection, present an overview of current research directions, and discuss open problems and possible future directions, all this with a focus on the classifiers and architectures of the detector, rather than on the used features.

2. A BRIEF REVIEW OF OBJECT DETECTION RESEARCH

Early works on object detection were based on template matching techniques and simple part-based models [e.g., Fischler and Elschlager (1973)]. Later, methods based on statistical classifiers (e.g., Neural Networks, SVM, Adaboost, Bayes, etc.) were introduced [e.g., Osuna et al. (1997), Rowley et al. (1998), Sung and Poggio (1998), Schneiderman and Kanade (2000), Yang et al. (2000a,b), Fleuret and Geman (2001), Romdhani et al. (2001), and Viola and Jones (2001)]. This initial successful family of object detectors, all of them based on statistical classifiers, set the ground for most of the following research in terms of training and evaluation procedures and classification techniques.

Because face detection is a critical ability for any system that interacts with humans, it is the most common application of object detection. However, many additional detection problems have been studied [e.g., Papageorgiou and Poggio (2000), Agarwal et al. (2004), Alexe et al. (2010), Everingham et al. (2010), and Andreopoulos and Tsotsos (2013)]. Most cases correspond to objects that people often interact with, such as other humans [e.g., pedestrians (Papageorgiou and Poggio, 2000; Viola and Jones, 2002; Dalal and Triggs, 2005; Bourdev et al., 2010; Paisitkriangkrai et al., 2015)] and body parts [(Kölsch and Turk, 2004; Ong and Bowden, 2004; Wu and Nevatia, 2005; Verschae et al., 2008; Bourdev and Malik, 2009) e.g., faces, hands, and eyes], as well as vehicles [(Papageorgiou and Poggio, 2000; Felzenszwalb et al., 2010b), e.g., cars and airplanes], and animals [e.g., Fleuret and Geman (2008)].

Most object detection systems consider the same basic scheme, commonly known as *sliding window*: in order to detect the objects appearing in the image at different scales and locations, an exhaustive search is applied. This search makes use of a classifier, the core part of the detector, which indicates if a given image patch, corresponds to the object or not. Given that the classifier basically works at a given scale and patch size, several versions of the input image are generated at different scales, and the classifier is used to classify all possible patches of the given size, for each of the downscaled versions of the image.

Basically, three alternatives exist to the sliding window scheme. The first one is based on the use of bag-of-words (Weinland et al., 2011; Tsai, 2012), method sometimes used for verifying the presence of the object, and that in some cases can be efficiently applied by iteratively refining the image region that contains the object [e.g., Lampert et al. (2009)]. The second one samples patches and iteratively searches for regions of the image where it is likely that the object is present [e.g., Prati et al. (2012)]. These two schemes reduce the number of image patches where to perform the classification, seeking to avoid an exhaustive search over all image patches. The third scheme finds key-points and then matches them to perform the detection [e.g., Azzopardi and Petkov (2013)]. These schemes cannot always guarantee that all object's instances will be detected.

3. OBJECT DETECTION APPROACHES

Object detection methods can be grouped in five categories, each with merits and demerits: while some are more robust, others can be used in real-time systems, and others can be handle more classes, etc. **Table 1** gives a qualitative comparison.

3.1. Coarse-to-Fine and Boosted Classifiers

The most popular work in this category is the boosted cascade classifier of Viola and Jones (2004). It works by efficiently rejecting, in a cascade of test/filters, image patches that do not correspond to the object. Cascade methods are commonly used with boosted classifiers due to two main reasons: (i) boosting generates an additive classifier, thus it is easy to control the complexity of each stage of the cascade and (ii) during training, boosting can be also used for feature selection, allowing the use of large (parametrized) families of features. A coarse-to-fine cascade classifier is usually the first kind of classifier to consider when efficiency is a key requirement. Recent methods based on boosted classifiers include Li and Zhang (2004), Gangaputra and Geman (2006), Huang et al. (2007), Wu and Nevatia (2007), Verschae et al. (2008), and Verschae and Ruiz-del-Solar (2012).

3.2. Dictionary Based

The best example in this category is the Bag of Word method [e.g., Serre et al. (2005) and Mutch and Lowe (2008)]. This approach is basically designed to detect a single object per image, but after removing a detected object, the remaining objects can be detected [e.g., Lampert et al. (2009)]. Two problems with this approach are that it cannot robustly handle well the case of two instances of the object appearing near each other, and that the localization of the object may not be accurate.

3.3. Deformable Part-Based Model

This approach considers object and part models and their relative positions. In general, it is more robust that other approaches, but it is rather time consuming and cannot detect objects appearing at small scales. It can be traced back to the deformable models (Fischler and Elschlager, 1973), but successful methods are recent (Felzenszwalb et al., 2010b). Relevant works include Felzenszwalb et al. (2010a) and Yan et al. (2014), where efficient evaluation of deformable part-based model is implemented using a coarse-to-fine cascade model for faster evaluation, Divvala et al. (2012), where the relevance of the part-models is analyzed, among others [e.g., Azizpour and Laptev (2012), Zhu and Ramanan (2012), and Girshick et al. (2014)].

3.4. Deep Learning

One of the first successful methods in this family is based on convolutional neural networks (Delakis and Garcia, 2004). The key difference between this and the above approaches is that in this approach the feature representation is learned instead of being designed by the user, but with the drawback that a large number of training samples is required for training the classifier. Recent methods include Dean et al. (2013), Huval et al. (2013), Ouyang and Wang (2013), Sermanet et al. (2013), Szegedy et al. (2013), Zeng et al. (2013), Erhan et al. (2014), Zhou et al. (2014), and Ouyang et al. (2015).

3.5. Trainable Image Processing Architectures

In such architectures, the parameters of predefined operators and the combination of the operators are learned, sometimes considering an abstract notion of fitness. These are general-purpose architectures, and thus they can be used to build several modules of a larger system (e.g., object recognition, key point detectors and object detection modules of a robot vision system). Examples include trainable COSFIRE filters (Azzopardi and Petkov, 2013, 2014), and Cartesian Genetic Programming (CGP) (Harding et al., 2013; Leitner et al., 2013).

4. CURRENT RESEARCH PROBLEMS

Table 2 presents a summary of solved, current, and open problems. In the present section we discuss current research directions.

4.1. Multi-Class

Many applications require detecting more than one object class. If a large number of classes is being detected, the processing speed becomes an important issue, as well as the kind of classes that the system can handle without accuracy loss. Works that have addressed the multi-class detection problem include Torralba et al. (2007), Razavi et al. (2011), Benbouzid et al. (2012),

TABLE 1 | Qualitative comparison of object detection approaches.

Method	Coarse-to-fine and boosted classifiers	Dictionary based	Deformable part-based models	Deep learning	Trainable image processing architectures
Accuracy	++	+=	++	++	+=
Generality	==	++	+=	++	+=
Speed	++	+=	==	+=	+=
Advantages	Real-time, it can work at small resolutions	Representation can be shared across classes	It can handle deformations and occlusions	Representation can be transfered to other classes	General-purpose architecture that can be used is several modules of a system
Drawbacks/ requirements	Features are predefined	It may not detect all object instances	It can not detect small objects	Large training sets specialized hardware (GPU) for efficiency	The obtained system may be Too specialized for a particular setting
Typical applications	Robotics, security	Retrieval, search	Transportation pedestrian detection	Retrieval, search	HCI, health, robotics

Accuracy: ++, High; +=, Good; ==, Low.
Speed: ++, real-time (15 fps or more); +=, online (10–5 fps); ==, offline (5 fps or more).
Generality: ++ (+=), applicable to many (some) object classes; ==, depend on features designed for specific classes.

TABLE 2 | Summary of current directions and open problems.

Solved problems	Single-class	Single-view	Small deformations	Multi-scale
Current directions	Multi-class (scalability and efficiency)	Multi-view/pose Multi-resolution	Occlusions, deformable Interlaced object and background	Contextual information Temporal features
Open	Incremental learning	Object-part relation	Pixel-level detection Background objects	Multi-modal

Song et al. (2012), Verschae and Ruiz-del-Solar (2012), and Erhan et al. (2014). Efficiency has been addressed, e.g., by using the same representation for several object classes, as well as by developing multi-class classifiers designed specifically to detect multiple classes. Dean et al. (2013) presents one of the few existing works for very large-scale multi-class object detection, where 100,000 object classes were considered.

4.2. Multi-View, Multi-Pose, Multi-Resolution

Most methods used in practice have been designed to detect a single object class under a single view, thus these methods cannot handle multiple views, or large pose variations; with the exception of deformable part-based models which can deal with some pose variations. Some works have tried to detect objects by learning subclasses (Wu and Nevatia, 2007) or by considering views/poses as different classes (Verschae and Ruiz-del-Solar, 2012); in both cases improving the efficiency and robustness. Also, multi-pose models [e.g., Erol et al. (2007)] and multi-resolution models [e.g., Park et al. (2010)] have been developed.

4.3. Efficiency and Computational Power

Efficiency is an issue to be taken into account in any object detection system. As mentioned, a coarse-to-fine classifier is usually the first kind of classifier to consider when efficiency is a key requirement [e.g., Viola et al. (2005)], while reducing the number of image patches where to perform the classification [e.g., Lampert et al. (2009)] and efficiently detecting multiple classes [e.g., Verschae and Ruiz-del-Solar (2012)] have also been used. Efficiency does not imply real-time performance, and works such as Felzenszwalb et al. (2010b) are robust and efficient, but not fast enough for real-time problems. However, using specialized hardware (e.g., GPU) some methods can run in real-time (e.g., deep learning).

4.4. Occlusions, Deformable Objects, and Interlaced Object and Background

Dealing with partial occlusions is also an important problem, and no compelling solution exits, although relevant research has been done [e.g., Wu and Nevatia (2005)]. Similarly, detecting objects that are not "closed," i.e., where objects and background pixels are interlaced with background is still a difficult problem. Two examples are hand detection [e.g., Kölsch and Turk (2004)] and pedestrian detection [see Dollar et al. (2012)]. Deformable part-based model [e.g., Felzenszwalb et al. (2010b)] have been to some extend successful under this kind of problem, but further improvement is still required.

4.5. Contextual Information and Temporal Features

Integrating contextual information (e.g., about the type of scene, or the presence of other objects) can increase speed and robustness, but "when and how" to do this (before, during or after the detection), it is still an open problem. Some proposed solutions include the use of (i) spatio-temporal context [e.g., Palma-Amestoy et al. (2010)], (ii) spatial structure among visual words [e.g., Wu et al. (2009)], and (iii) semantic information aiming to map semantically related features to visual words [e.g., Wu et al. (2010)], among many others [e.g., Torralba and Sinha (2001), Divvala et al. (2009), Sun et al. (2012), Mottaghi et al. (2014), and Cadena et al. (2015)]. While most methods consider the detection of objects in a single frame, temporal features can be beneficial [e.g., Viola et al. (2005) and Dalal et al. (2006)].

5. OPEN PROBLEMS AND FUTURE DIRECTIONS

In the following, we outline problems that we believe have not been addressed, or addressed only partially, and may be interesting relevant research directions.

5.1. Open-World Learning and Active Vision

An important problem is to incrementally learn, to detect new classes, or to incrementally learn to distinguish among subclasses after the "main" class has been learned. If this can be done in an unsupervised way, we will be able to build new classifiers based on existing ones, without much additional effort, greatly reducing the effort required to learn new object classes. Note that humans are continuously inventing new objects, fashion changes, etc., and therefore detection systems will need to be continuously updated, adding new classes, or updating existing ones. Some recent works have addressed these issues, mostly based on deep learning and transfer learning methods [e.g., Bengio (2012), Mesnil et al. (2012), and Kotzias et al. (2014)]. This open-world learning is of particular importance in robot applications, case where active vision mechanisms can aid in the detection and learning [e.g., Paletta and Pinz (2000) and Correa et al. (2012)].

5.2. Object-Part Relation

During the detection process, should we detect the object first or the parts first? This is a basic dilemma, and no clear solution exists. Probably, the search for the object and for the parts must be done concurrently where both processes give feedback to each other. How to do this is still an open problem and is likely related to how to use of context information. Moreover, in cases the object part can be also decomposed in subparts, an interaction among several hierarchies emerge, and in general it is not clear what should be done first.

5.3. Multi-Modal Detection

The use of new sensing modalities, in particular depth and thermal cameras, has seen some development in the last years [e.g., Fehr and Burkhardt (2008) and Correa et al. (2012)]. However, the methods used for processing visual images are also used for thermal images, and to a lesser degree for depth images. While using thermal images makes easier to discriminate the foreground from the background, it can only be applied to objects that irradiate infrared light (e.g., mammals, heating, etc.). Using depth images is easy to segment the objects, but general methods for detecting specific classes has not been proposed, and probably higher resolution depth images are required. It seems that depth and thermal cameras alone are not enough for object detection, at least with their current resolution, but further advances can be expected as the sensing technology improves.

5.4. Pixel-Level Detection (Segmentation) and Background Objects

In many applications, we may be interested in detecting objects that are usually considered as background. The detection of such "background objects," such as rivers, walls, mountains, has not been addressed by most of the here mentioned approaches. In general, this kind of problem has been addressed by first segmenting the image and later labeling each segment of the image [e.g., Peng et al. (2013)]. Of course, for successfully detecting all objects in a scene, and to completely understand the scene, we will need to have a pixel level detection of the objects, and further more, a 3D model of such scene. Therefore, at some point object detection and image segmentation methods may need to be integrated. We are still far from attaining such automatic understanding of the world, and to achieve this, active vision mechanisms might be required [e.g., Aloimonos et al. (1988) and Cadena et al. (2015)].

6. CONCLUSION

Object detection is a key ability for most computer and robot vision system. Although great progress has been observed in the last years, and some existing techniques are now part of many consumer electronics (e.g., face detection for auto-focus in smartphones) or have been integrated in assistant driving technologies, we are still far from achieving human-level performance, in particular in terms of open-world learning. It should be noted that object detection has not been used much in many areas where it could be of great help. As mobile robots, and in general autonomous machines, are starting to be more widely deployed (e.g., quad-copters, drones and soon service robots), the need of object detection systems is gaining more importance. Finally, we need to consider that we will need object detection systems for nano-robots or for robots that will explore areas that have not been seen by humans, such as depth parts of the sea or other planets, and the detection systems will have to learn to new object classes as they are encountered. In such cases, a real-time open-world learning ability will be critical.

ACKNOWLEDGMENTS

This research was partially funded by the FONDECYT Projects 3120218 and 1130153 (CONICYT, Chile).

REFERENCES

Agarwal, S., Awan, A., and Roth, D. (2004). Learning to detect objects in images via a sparse, part-based representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 26, 1475–1490. doi:10.1109/TPAMI.2004.108

Alexe, B., Deselaers, T., and Ferrari, V. (2010). "What is an object?," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* (San Francisco, CA: IEEE), 73–80. doi:10.1109/CVPR.2010.5540226

Aloimonos, J., Weiss, I., and Bandyopadhyay, A. (1988). Active vision. *Int. J. Comput. Vis.* 1, 333–356. doi:10.1007/BF00133571

Andreopoulos, A., and Tsotsos, J. K. (2013). 50 years of object recognition: directions forward. *Comput. Vis. Image Underst.* 117, 827–891. doi:10.1016/j.cviu.2013.04.005

Azizpour, H., and Laptev, I. (2012). "Object detection using strongly-supervised deformable part models," in *Computer Vision-ECCV 2012* (Florence: Springer), 836–849.

Azzopardi, G., and Petkov, N. (2013). Trainable cosfire filters for keypoint detection and pattern recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 490–503. doi:10.1109/TPAMI.2012.106

Azzopardi, G., and Petkov, N. (2014). Ventral-stream-like shape representation: from pixel intensity values to trainable object-selective cosfire models. *Front. Comput. Neurosci.* 8:80. doi:10.3389/fncom.2014.00080

Benbouzid, D., Busa-Fekete, R., and Kegl, B. (2012). "Fast classification using sparse decision dags," in *Proceedings of the 29th International Conference on Machine Learning (ICML-12), ICML '12*, eds J. Langford and J. Pineau (New York, NY: Omnipress), 951–958.

Bengio, Y. (2012). "Deep learning of representations for unsupervised and transfer learning," in *ICML Unsupervised and Transfer Learning, Volume 27 of JMLR Proceedings*, eds I. Guyon, G. Dror, V. Lemaire, G. W. Taylor, and D. L. Silver (Bellevue: JMLR.Org), 17–36.

Bourdev, L. D., Maji, S., Brox, T., and Malik, J. (2010). "Detecting people using mutually consistent poselet activations," in *Computer Vision – ECCV 2010 – 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part VI, Volume 6316 of Lecture Notes in Computer Science*, eds K. Daniilidis, P. Maragos, and N. Paragios (Heraklion: Springer), 168–181.

Bourdev, L. D., and Malik, J. (2009). "Poselets: body part detectors trained using 3d human pose annotations," in *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 – October 4, 2009* (Kyoto: IEEE), 1365–1372.

Cadena, C., Dick, A., and Reid, I. (2015). "A fast, modular scene understanding system using context-aware object detection," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on* (Seattle, WA).

Correa, M., Hermosilla, G., Verschae, R., and Ruiz-del-Solar, J. (2012). Human detection and identification by robots using thermal and visual information in domestic environments. *J. Intell. Robot Syst.* 66, 223–243. doi:10.1007/s10846-011-9612-2

Dalal, N., and Triggs, B. (2005). "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 1 (San Diego, CA: IEEE), 886–893. doi:10.1109/CVPR.2005.177

Dalal, N., Triggs, B., and Schmid, C. (2006). "Human detection using oriented histograms of flow and appearance," in *Computer Vision ECCV 2006, Volume 3952 of Lecture Notes in Computer Science*, eds A. Leonardis, H. Bischof, and A. Pinz (Berlin: Springer), 428–441.

Dean, T., Ruzon, M., Segal, M., Shlens, J., Vijayanarasimhan, S., Yagnik, J., et al. (2013). "Fast, accurate detection of 100,000 object classes on a single machine," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on* (Washington, DC: IEEE), 1814–1821.

Delakis, M., and Garcia, C. (2004). Convolutional face finder: a neural architecture for fast and robust face detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 26, 1408–1423. doi:10.1109/TPAMI.2004.97

Divvala, S., Hoiem, D., Hays, J., Efros, A., and Hebert, M. (2009). "An empirical study of context in object detection," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (Miami, FL: IEEE), 1271–1278. doi:10.1109/CVPR.2009.5206532

Divvala, S. K., Efros, A. A., and Hebert, M. (2012). "How important are deformable parts in the deformable parts model?," in *Computer Vision-ECCV 2012. Workshops and Demonstrations* (Florence: Springer), 31–40.

Dollar, P., Wojek, C., Schiele, B., and Perona, P. (2012). Pedestrian detection: an evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 743–761. doi:10.1109/TPAMI.2011.155

Endres, I., and Hoiem, D. (2010). "Category independent object proposals," in *Proceedings of the 11th European Conference on Computer Vision: Part V, ECCV'10* (Berlin: Springer-Verlag), 575–588.

Enzweiler, M., and Gavrila, D. (2009). Monocular pedestrian detection: survey and experiments. *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 2179–2195. doi:10.1109/TPAMI.2008.260

Erhan, D., Szegedy, C., Toshev, A., and Anguelov, D. (2014). "Scalable object detection using deep neural networks," in *Computer Vision and Pattern Recognition*

(CVPR), 2014 IEEE Conference on (Columbus, OH: IEEE), 2155–2162. doi:10.1109/CVPR.2014.276

Erol, A., Bebis, G., Nicolescu, M., Boyle, R. D., and Twombly, X. (2007). Vision-based hand pose estimation: a review. *Comput. Vis. Image Underst.* 108, 52–73; Special Issue on Vision for Human-Computer Interaction. doi:10.1016/j.cviu.2006.10.012

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* 88, 303–338. doi:10.1007/s11263-009-0275-4

Fehr, J., and Burkhardt, H. (2008). "3d rotation invariant local binary patterns," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on* (Tampa, FL: IEEE), 1–4. doi:10.1109/ICPR.2008.4761098

Felzenszwalb, P. F., Girshick, R. B., and McAllester, D. (2010a). "Cascade object detection with deformable part models," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* (San Francisco, CA: IEEE), 2241–2248.

Felzenszwalb, P., Girshick, R., McAllester, D., and Ramanan, D. (2010b). Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* 32, 1627–1645. doi:10.1109/TPAMI.2009.167

Fischler, M. A., and Elschlager, R. (1973). The representation and matching of pictorial structures. *IEEE Trans. Comput.* C-22, 67–92. doi:10.1109/T-C.1973.223602

Fleuret, F., and Geman, D. (2001). Coarse-to-fine face detection. *Int. J. Comput. Vis.* 41, 85–107. doi:10.1023/A:1011113216584

Fleuret, F., and Geman, D. (2008). Stationary features and cat detection. *Journal of Machine Learning Research (JMLR)* 9, 2549–2578.

Gangaputra, S., and Geman, D. (2006). "A design principle for coarse-to-fine classification," in *Proc. of the IEEE Conference of Computer Vision and Pattern Recognition*, Vol. 2 (New York, NY: IEEE), 1877–1884. doi:10.1109/CVPR.2006.21

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on* (Columbus, OH: IEEE), 580–587.

Harding, S., Leitner, J., and Schmidhuber, J. (2013). "Cartesian genetic programming for image processing," in *Genetic Programming Theory and Practice X, Genetic and Evolutionary Computation*, eds R. Riolo, E. Vladislavleva, M. D. Ritchie, and J. H. Moore (New York, NY: Springer), 31–44.

Hjelmås, E., and Low, B. K. (2001). Face detection: a survey. *Comput. Vis. Image Underst.* 83, 236–274. doi:10.1006/cviu.2001.0921

Huang, C., Ai, H., Li, Y., and Lao, S. (2007). High-performance rotation invariant multiview face detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 671–686. doi:10.1109/TPAMI.2007.1011

Huval, B., Coates, A., and Ng, A. (2013). *Deep Learning for Class-Generic Object Detection*. arXiv preprint arXiv:1312.6885.

Kölsch, M., and Turk, M. (2004). "Robust hand detection," in *Proceedings of the Sixth International Conference on Automatic Face and Gesture Recognition* (Seoul: IEEE), 614–619.

Kotzias, D., Denil, M., Blunsom, P., and de Freitas, N. (2014). *Deep Multi-Instance Transfer Learning*. CoRR, abs/1411.3128.

Lampert, C. H., Blaschko, M., and Hofmann, T. (2009). Efficient subwindow search: a branch and bound framework for object localization. *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 2129–2142. doi:10.1109/TPAMI.2009.144

Leitner, J., Harding, S., Chandrashekhariah, P., Frank, M., Frster, A., Triesch, J., et al. (2013). Learning visual object detection and localisation using icvision. *Biol. Inspired Cogn. Archit.* 5, 29–41; Extended versions of selected papers from the Third Annual Meeting of the {BICA} Society (BICA 2012). doi:10.1016/j.bica.2013.05.009

Li, J., and Allinson, N. M. (2008). A comprehensive review of current local features for computer vision. *Neurocomputing* 71, 1771–1787; Neurocomputing for Vision Research Advances in Blind Signal Processing. doi:10.1016/j.neucom.2007.11.032

Li, S. Z., and Zhang, Z. (2004). Floatboost learning and statistical face detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 26, 1112–1123. doi:10.1109/TPAMI.2004.68

Li, Y., Wang, S., Tian, Q., and Ding, X. (2015). Feature representation for statistical-learning-based object detection: a review. *Pattern Recognit.* 48, 3542–3559. doi:10.1016/j.patcog.2015.04.018

Mesnil, G., Dauphin, Y., Glorot, X., Rifai, S., Bengio, Y., Goodfellow, I. J., et al. (2012). "Unsupervised and transfer learning challenge: a deep learning approach," in *JMLR W& CP: Proceedings of the Unsupervised and Transfer Learning Challenge and Workshop*, Vol. 27, eds I. Guyon, G. Dror, V. Lemaire, G. Taylor, and D. Silver (Bellevue: JMLR.org) 97–110.

Mottaghi, R., Chen, X., Liu, X., Cho, N.-G., Lee, S.-W., Fidler, S., et al. (2014). "The role of context for object detection and semantic segmentation in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on* (Columbus, OH: IEEE), 891–898. doi:10.1109/CVPR.2014.119

Mutch, J., and Lowe, D. G. (2008). Object class recognition and localization using sparse features with limited receptive fields. *Int. J. Comput. Vis.* 80, 45–57. doi:10.1007/s11263-007-0118-0

Ong, E.-J., and Bowden, R. (2004). "A boosted classifier tree for hand shape detection," in *Proceedings of the Sixth International Conference on Automatic Face and Gesture Recognition* (Seoul: IEEE), 889–894. doi:10.1109/AFGR.2004.1301646

Osuna, E., Freund, R., and Girosi, F. (1997). "Training support vector machines: an application to face detection," in *Proc. of the IEEE Conference of Computer Vision and Pattern Recognition* (San Juan: IEEE), 130–136. doi:10.1109/CVPR.1997.609310

Ouyang, W., and Wang, X. (2013). "Joint deep learning for pedestrian detection," in *Computer Vision (ICCV), 2013 IEEE International Conference on* (Sydney, VIC: IEEE), 2056–2063. doi:10.1109/ICCV.2013.257

Ouyang, W., Wang, X., Zeng, X., Qiu, S., Luo, P., Tian, Y., et al. (2015). "Deepid-net: deformable deep convolutional neural networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA: IEEE), 2403–2412.

Paisitkriangkrai, S., Shen, C., and van den Hengel, A. (2015). Pedestrian detection with spatially pooled features and structured ensemble learning. *IEEE Trans. Pattern Anal. Mach. Intell.* PP, 1. doi:10.1109/TPAMI.2015.2474388

Paletta, L., and Pinz, A. (2000). Active object recognition by view integration and reinforcement learning. *Rob. Auton. Syst.* 31, 71–86. doi:10.1016/S0921-8890(99)00079-2

Palma-Amestoy, R., Ruiz-del Solar, J., Yanez, J. M., and Guerrero, P. (2010). Spatiotemporal context integration in robot vision. *Int. J. Human. Robot.* 07, 357–377. doi:10.1142/S0219843610002192

Papageorgiou, C., and Poggio, T. (2000). A trainable system for object detection. *Int. J. Comput. Vis.* 38, 15–33. doi:10.1023/A:1008162616689

Park, D., Ramanan, D., and Fowlkes, C. (2010). "Multiresolution models for object detection," in *Computer Vision ECCV 2010, Volume 6314 of Lecture Notes in Computer Science*, eds K. Daniilidis, P. Maragos, and N. Paragios (Berlin: Springer), 241–254.

Peng, B., Zhang, L., and Zhang, D. (2013). A survey of graph theoretical approaches to image segmentation. *Pattern Recognit.* 46, 1020–1038. doi:10.1016/j.patcog.2012.09.015

Prati, A., Gualdi, G., and Cucchiara, R. (2012). Multistage particle windows for fast and accurate object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 1589–1604. doi:10.1109/TPAMI.2011.247

Ramanan, A., and Niranjan, M. (2012). A review of codebook models in patch-based visual object recognition. *J. Signal Process. Syst.* 68, 333–352. doi:10.1007/s11265-011-0622-x

Razavi, N., Gall, J., and Van Gool, L. (2011). "Scalable multi-class object detection," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (Providence, RI: IEEE), 1505–1512. doi:10.1109/CVPR.2011.5995441

Romdhani, S., Torr, P., Scholkopf, B., and Blake, A. (2001). "Computationally efficient face detection," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, Vol. 2 (Vancouver, BC: IEEE), 695–700. doi:10.1109/ICCV.2001.937694

Rowley, H. A., Baluja, S., and Kanade, T. (1998). Neural network-based detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 23–28. doi:10.1109/34.655647

Schneiderman, H., and Kanade, T. (2000). "A statistical model for 3D object detection applied to faces and cars," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* (Hilton Head, SC: IEEE), 746–751.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. (2013). *Overfeat: Integrated Recognition, Localization and Detection Using Convolutional Networks*. arXiv preprint arXiv:1312.6229.

Serre, T., Wolf, L., and Poggio, T. (2005). "Object recognition with features inspired by visual cortex," in *CVPR (2)* (San Diego, CA: IEEE Computer Society), 994–1000.

Song, H. O., Zickler, S., Althoff, T., Girshick, R., Fritz, M., Geyer, C., et al. (2012). "Sparselet models for efficient multiclass object detection," in *Computer Vision-ECCV 2012* (Florence: Springer), 802–815.

Sun, M., Bao, S., and Savarese, S. (2012). Object detection using geometrical context feedback. *Int. J. Comput. Vis.* 100, 154–169. doi:10.1007/s11263-012-0547-2

Sun, Z., Bebis, G., and Miller, R. (2006). On-road vehicle detection: a review. *IEEE Trans. Pattern Anal. Mach. Intell.* 28, 694–711. doi:10.1109/TPAMI.2006.104

Sung, K.-K., and Poggio, T. (1998). Example-based learning for viewed-based human face detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 39–51. doi:10.1109/34.655648

Szegedy, C., Toshev, A., and Erhan, D. (2013). "Deep neural networks for object detection," in *Advances in Neural Information Processing Systems 26*, eds C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (Harrahs and Harveys: Curran Associates, Inc), 2553–2561.

Torralba, A., Murphy, K. P., and Freeman, W. T. (2007). Sharing visual features for multiclass and multiview object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 854–869. doi:10.1109/TPAMI.2007.1055

Torralba, A., and Sinha, P. (2001). "Statistical context priming for object detection," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, Vol. 1 (Vancouver, BC: IEEE), 763–770. doi:10.1109/ICCV.2001.937604

Tsai, C.-F. (2012). Bag-of-words representation in image annotation: a review. *ISRN Artif. Intell.* 2012, 19. doi:10.5402/2012/376804

Tuytelaars, T., and Mikolajczyk, K. (2008). Local invariant feature detectors: a survey. *Found. Trends Comput. Graph. Vis.* 3, 177–280. doi:10.1561/0600000017

Verschae, R., and Ruiz-del-Solar, J. (2012). "Tcas: a multiclass object detector for robot and computer vision applications," in *Advances in Visual Computing, Volume 7431 of Lecture Notes in Computer Science*, eds G. Bebis, R. Boyle, B. Parvin, D. Koracin, C. Fowlkes, S. Wang, et al. (Berlin: Springer), 632–641.

Verschae, R., Ruiz-del-Solar, J., and Correa, M. (2008). A unified learning framework for object detection and classification using nested cascades of boosted classifiers. *Mach. Vis. Appl.* 19, 85–103. doi:10.1007/s00138-007-0084-0

Viola, P., and Jones, M. (2001). "Rapid object detection using a boosted cascade of simple features," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* (Kauai: IEEE), 511–518. doi:10.1109/CVPR.2001.990517

Viola, P., and Jones, M. (2002). "Fast and robust classification using asymmetric adaboost and a detector cascade," in *Advances in Neural Information Processing System 14* (Vancouver: MIT Press), 1311–1318.

Viola, P., Jones, M., and Snow, D. (2005). Detecting pedestrians using patterns of motion and appearance. *Int. J. Comput. Vis.* 63, 153–161. doi:10.1007/s11263-005-6644-8

Viola, P., and Jones, M. J. (2004). Robust real-time face detection. *Int. J. Comput. Vis.* 57, 137–154. doi:10.1023/B:VISI.0000013087.49260.fb

Weinland, D., Ronfard, R., and Boyer, E. (2011). A survey of vision-based methods for action representation, segmentation and recognition. *Comput. Vis. Image Underst.* 115, 224–241. doi:10.1016/j.cviu.2010.10.002

Wu, B., and Nevatia, R. (2005). "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in *ICCV '05: Proceedings of the 10th IEEE Int. Conf. on Computer Vision (ICCV'05) Vol 1* (Washington, DC: IEEE Computer Society), 90–97.

Wu, B., and Nevatia, R. (2007). "Cluster boosted tree classifier for multi-view, multi-pose object detection," in *ICCV* (Rio de Janeiro: IEEE), 1–8.

Wu, L., Hoi, S., and Yu, N. (2010). Semantics-preserving bag-of-words models and applications. *IEEE Trans. Image Process.* 19, 1908–1920. doi:10.1109/TIP.2010.2045169

Wu, L., Hu, Y., Li, M., Yu, N., and Hua, X.-S. (2009). Scale-invariant visual language modeling for object categorization. *IEEE Trans. Multimedia* 11, 286–294. doi:10.1109/TMM.2008.2009692

Yan, J., Lei, Z., Wen, L., and Li, S. Z. (2014). "The fastest deformable part model for object detection," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on* (Columbus, OH: IEEE), 2497–2504.

Yang, M.-H., Ahuja, N., and Kriegman, D. (2000a). "Mixtures of linear subspaces for face detection," in *Proc. Fourth IEEE Int. Conf. on Automatic Face and Gesture Recognition* (Grenoble: IEEE), 70–76.

Yang, M.-H., Roth, D., and Ahuja, N. (2000b). "A SNoW-based face detector," in *Advances in Neural Information Processing Systems 12* (Denver: MIT press), 855–861.

Yang, M.-H., Kriegman, D., and Ahuja, N. (2002). Detecting faces in images: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 24, 34–58. doi:10.1109/34.982883

Zafeiriou, S., Zhang, C., and Zhang, Z. (2015). A survey on face detection in the wild: past, present and future. *Comput. Vis. Image Underst.* 138, 1–24. doi:10.1016/j.cviu.2015.03.015

Zeng, X., Ouyang, W., and Wang, X. (2013). "Multi-stage contextual deep learning for pedestrian detection," in *Computer Vision (ICCV), 2013 IEEE International Conference on* (Washington, DC: IEEE), 121–128.

Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., and Torralba, A. (2014). *Object Detectors Emerge in Deep Scene Cnns*. CoRR, abs/1412.6856.

Zhu, X., and Ramanan, D. (2012). "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (Providence: IEEE), 2879–2886.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.