Electronics Letters 51.23 (2015): 1865 – 1867

DOI: http://dx.doi.org/10.1049/el.2015.3099

Context-aware part-based people detection for video monitoring

Alvaro García-Martín and Juan C. SanMiguel

We propose a novel approach for part-based people detection in images that uses contextual information. Two sources of context are distinguished regarding the local (neighbour) information and the relative importance of the parts in the model. Local context determines part visibility which is derived from the spatial location of static objects in the scene and from the relation between scales of analysis and detection window sizes. Experimental results over various datasets show that the proposed use of context outperforms the related state-of-the-art.

Introduction: People detection in images faces many challenges related to pose changes, illumination variations, occlusions and clutter. Holistic and part-based models have been proposed to cope with such high complexity and variability. *Holistic models* describe a person as a whole by means of region, shape or colour features such as Histograms of Oriented Gradients [1] and Aggregated Channel Features (ACF) [2]. *Deformable Part Models* (DPMs) [3] consider a person as a root component and P body parts, thus providing a superior ability to handle variations in the relative locations of parts. Recent DPM improvements have targeted partial occlusions which still presents a major hurdle. Occlusion patterns can be included in the model by learning from annotated training data [4] or by means of double-person detectors [5]. However, both approaches re-train DPMs which is time consuming. Moreover, non-visibility of parts can be assumed to create models based on subsets of parts [6], resulting in a large set of configurations. Although recent efforts address partial occlusions, the use of context in DPMs has received less attention. For example, contextual cues are obtained from nearby background pixels to get a holistic classifier [7]. Region segmentation can help to filter out wrong hypothesised part locations [8]. Additional deformation models and AND-OR combinations are acquired from specific training data to adapt DPMs to a particular context [9]. Albeit effective, these context-based DPMs require further training which makes not straightforward their adaptation to other contexts.

To overcome these shortcomings, this Letter presents a framework for people detection based on context that extends DPMs for occluded object detection. This proposal employs two sources of context for the local (neighbour) information and the relative part importance. We simplify the acquisition and use of context so the proposed approach can be easily adapted to new contexts without requiring additional training.

Framework for context-aware people detection: We include contextual information in DPMs [3]. Each part p is represented by a 3-tuple $\{F_p, D_p, v_p\}$ where F_p is the appearance model, D_p is the deformation model and v_p is the optimum location of the part. Detecting people in a $M \times N$ image I involves computing a score s for hypothesised locations of all parts, defined as $\{l_0, ...l_P\}$ where l_p is a spatial position (x, y) and analysis scale a. We extend DPMs to use the context of each hypothesis via contextual part scores $\Upsilon(l_p, \mathcal{C}_p)$ where $\mathcal{C}_p$ is the scene knowledge of part p. The score s for each hypothesis is computed as:

$$s(l_{0,...,}l_P) = \sum_{p=0}^{P} \Upsilon(l_p, \mathcal{C}_p) \Big[\langle F_p, \phi(l_p, I) \rangle + \langle D_p, \psi(l_0, l_p) \rangle \Big] \quad (1)$$

where $\langle \cdot, \cdot \rangle$ is the scalar product, $\phi(l_p, I)$ are the image features from I at location l_p and $\psi(l_0, l_p)$ is a 4D descriptor for the displacements between the hypothesised l_p and optimum v_p part locations with respect to the root location l_0. For each part, the contextual score $\Upsilon(l_p, \mathcal{C}_p)$ is decomposed into the local context φ^l and the relative context φ^r as:

$$\Upsilon(l_p, \mathcal{C}_p) = \varphi^r \Big(\varphi^l(l_p, \mathcal{C}_p), \{\varphi^l(l_0, \mathcal{C}_0), ..., \varphi^l(l_P, \mathcal{C}_P)\} \setminus \{\varphi^l(l_p, \mathcal{C}_p)\} \Big) \quad (2)$$

where the local context φ^l refers to the spatial neighbourhood of the part using the knowledge $\mathcal{C}_p$. The relative context φ^r measures the importance of the local context for each part $\varphi^l(l_p, \mathcal{C}_p)$ as compared to the local context of the other parts $\{\varphi^l(l_0, \mathcal{C}_0), ..., \varphi^l(l_P, \mathcal{C}_P)\} \setminus \{\varphi^l(l_p, \mathcal{C}_p)\}$. DPM [3] defines an homogeneous part combination and context in (1) introduces part heterogeneity, which modifies the scores and the original detection thresholds cannot be used. The role of φ^r is to calibrate the context-based part combination to keep the original thresholds. For example, relative part importance can be derived as the Kullback-Leibler divergence between the score distributions of the full and one-part-out models [6]. Estimating φ^r does not require additional training since score distributions can be collected from the original training set. In this Letter, we focus on defining the local context $\varphi^l(l_p, \mathcal{C}_p)$ and the knowledge $\mathcal{C}_p$.

Local context for people detection: We consider two local contexts for $\varphi^l(l_p, \mathcal{C}_p)$ that explore spatial neighbourhood to determine parts visibility and, therefore, their importance when combined in DPMs. First, we define context according to the detection scale a. Parts of the model may fall outside of the image I at certain locations and scales, thus decreasing detection performance as these parts are not visible. To obtain the *scale context* $\varphi^l_a(l_p, \mathcal{C}_p) \equiv \varphi^l_{p,a}(x, y)$ for each part location (x, y) and analysis scale a, we apply a kernel K^a_p over an all-ones $M \times N$ matrix I':

$$\varphi^l_{p,a}(x, y) = \sum_{i=x}^{x+M'} \sum_{j=y}^{y+N'} I'(i, j) \cdot K^a_p(i + d_x - x, j + d_y - y) \quad (3)$$

where (i, j) are pixel coordinates, (d_x, d_y) are the part displacements with respect to the root center location and K^a_p is an all-ones $M' \times N'$ matrix whose size is the one of the rescaled part appearance model F_p by the factor a. $\varphi^l_{p,a}(x, y)$ estimates the likelihood to detect the part p in the image I at scale a and position (x, y). Hence, the aggregated scale context for all parts is defined as the sizes of the kernels at the considered scales, $\mathcal{C}_p = \{M' \times N'\}_{a=0...A}$. Fig. 1 depicts examples of the scale context.

a b c

Fig. 1 *Examples of scale context $\varphi^l_{p,a}$ for an 352×288 image using a= 2 (twice the original scale [3]). The values range from 1 (white) to 0 (black).*
a Root body part
b Head body part
c Shoulder-left body part

Second, we also estimate local context from domain knowledge descriptions such as the static scene objects [10], which are combined with spatial constraints into semantic rules in an ontology framework [10]. For example, some detections may be avoided such as for legs in the ceiling of a scene, heads in the floor of a scene or body parts occluded by tables. If we assume that this view-dependent context does not change over time, it can applied in video monitoring with static cameras. Otherwise, context needs to be updated accordingly. To obtain the *scene context* $\varphi^l_s(l_p, \mathcal{C}_p) \equiv \varphi^l_{p,s}(x, y)$ for each part p and scale a, we apply the previously defined kernel K^a_p similarly to (3):

$$\varphi^l_{p,s}(x, y) = \min_{\forall o \in \mathcal{O}_p} \left(\sum_{i=x}^{x+M'} \sum_{j=y}^{y+N'} \mathcal{M}_o(i, j) \cdot K^a_p(i + d_x - x, j + d_y - y) \right) \quad (4)$$

where $\mathcal{O}_p$ is the set of static objects o that may occlude (e.g. tables) or prevent (e.g. ceilings) the detection for the part p; $|\cdot|$ is the set cardinality and $\mathcal{M}_o$ is a binary $M \times N$ matrix indicating the location of the object o which can be obtained via annotation tools [10]. The semantic rules are represented by the sets $\mathcal{O}_p$, linking each part with the static objects affecting its visibility and, therefore, determining the part knowledge as $\mathcal{C}_p = \{\mathcal{O}_p, \mathcal{M}_o\}$. Fig. 2 depicts examples of the scene context.

a b c

Fig. 2 *Examples of scene context $\varphi^l_{p,s}$ for EDds dataset [12], using twice the original scale [3]. For the part maps, values range from 1 (white) to 0 (black).*
a Root body part
b Head body part
c The annotation of all stationary scene objects (each one as a unique colour)

Experimental setup: We test the proposed approach in datasets for video monitoring with static cameras: LIRIS [11], EDds [12] and PDbm [13]. These datasets provide heterogeneous test conditions covering common detection problems such as scale changes, occlusions and clutter. In total, 51075 people are manually annotated in 46917 frames. For the proposed approach, we use the original DPM [3] (`voc-release 4.1`) and distinguish three context-based derivations: DPM-A using $\varphi^l_{p,a}(l_p, \mathcal{C}_p)$ for scale, DPM-S using $\varphi^l_{p,s}(l_p, \mathcal{C}_p)$ for scene and DPM-B for the combination of both contexts as $\varphi^l_p = min(\varphi^l_{p,a}, \varphi^l_{p,s})$. We use the ontology framework [10] to provide the object context via the annotated spatial location of static objects. For the relative importance context φ^r in (1), we use the part weights as defined in [6]. Finally, we measure performance using the area under precision-recall (AUC-PR) curves [13] and provide comparisons against HOG [1], ACF-I [2] (using INRIA model), ACF-C [2] (using Caltech model) and DPM [3] approaches.

Experimental results: Table 1 compares the mean results obtained for each dataset. HOG is the worst due to the use of gray-scale holistic features such as oriented gradients whereas complex holistic approaches (ACF-I and ACF-C) significantly improve performance since multiple features are combined. DPM is the best selected approach which is outperformed by the proposal in all datasets, demonstrating that context is useful to increase detection performance in a variety of situations.

Table 1: Detection results for each dataset in terms of AUC-PR. $\%\Delta$ is the percentage increase of DPM-B against the best approach.

Dataset	HOG [1]	ACF-I [2]	ACF-C [2]	DPM [3]	DPM-B	$\%\Delta$
LIRIS	46.9	66.9	59.5	67.2	**86.1**	28.1
EDds	83.5	93.8	73.8	94.4	**98.3**	4.1
PDbm	48.2	73.4	60.5	75.1	**77.6**	3.3
Mean	59.5	78.0	64.6	78.9	**87.3**	10

Fig. 3 shows examples comparing the proposed and original approach. For clarity, we only show results for the best compared approach (DPM). For PDbm, no results are reported for DPM-S since no annotation is available for the scene objects. In all cases, DPM-B improves results thanks to the use of contextual information as seen in the PR curves. In Fig. 3(a), it is clear the use of both context-based derivations for the person on the left and only scene context (the table) for the person on the right. In Fig. 3(b), it can be observed the benefit of using scene context on both people where the table and ceiling scene objects help to detect the occluded people and avoid false positives, respectively. In Fig. 3(c), scale context increases detection performance for the person at the bottom of the image, demonstrative the robustness of DPM-B for detection scales where the person falls outside the image.

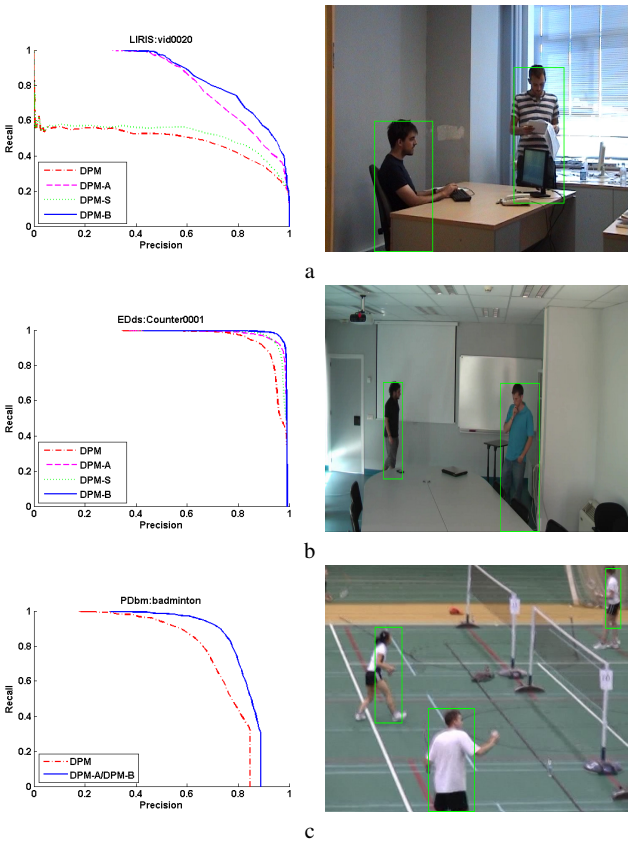

Fig. 3 *Comparative results between selected and proposed approaches using PR curves (left) and detected bounding boxes by DPM-B (right)*
a Frame 18 of *vid0020* sequence (LIRIS dataset)
b Frame 1238 of *Counter0001* sequence (EDds dataset)
c Frame 19 of *badminton* sequence (PDbm dataset)

Conclusions: This letter has presented a context-based DPM approach for people detection. The context is defined as the relative part importance in the model and the local (neighbourhood) information of hypothesised part locations. Local context is further explored to account for non-visible parts due to scale constraints and for occlusions due to scene objects, which can be easily provided as manual annotations. The proposed approach does not require re-training unlike related literature and the context is used to adapt the DPM combination. Performance increase over various datasets demonstrates the utility of context for people detection.

Acknowledgment: This work was supported by the Spanish Government (HA-Video TEC2014-5317-R).

A. García-Martín and J. C. SanMiguel (*Video Processing and Understanding Lab, Universidad Autónoma de Madrid, Madrid, Spain*)

E-mail: alvaro.garcia@uam.es

References

1 Dalal, N., Triggs, B.: 'Histograms of oriented gradients for human detection'. *IEEE Conf. Comput. Vision Patt. Recog.*, 2005, pp. 886-893
2 Dollár, P., Appel, R. and Kienzle, W.: 'Crosstalk Cascades for Frame-Rate Pedestrian Detection', *Eur. Conf. on Comput. Vision*, 2012, pp. 645-659.
3 Felzenszwalb, P., Girshick, R., McAllester, D., and Ramanan, D.: 'Object detection with discriminatively trained part-based models', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010, **32**, (9), pp. 1627-1645
4 Pepik, B., Stark, M., Gehler, P. and Schiele, B.: 'Occlusion Patterns for Object Class Detection', *IEEE Conf. Comput. Vision Patt. Recog.*, 2013, pp. 3284-3293.
5 Tang, S., Andriluka, M. and Schiele, B.: 'Detection and Tracking of Occluded People', *Int. J. Comput. Vision*, 2014, **110**, (1), pp. 58-69
6 García-Martín, A., Heras, R., and Sikora, T.: 'A Multi-configuration Part-based Person Detector', *Int. Conf. on Signal Processing and Multimedia Applications*, 2014, pp. 321-328.
7 Ding, Y., Xiao, J.: 'Contextual boost for pedestrian detection', *IEEE Conf. Comput. Vision Patt. Recog.*, 2012, pp. 2895-2902.
8 Trulls, E., Tsogkas, S., Kokkinos, I., Sanfeliu, A. and Moreno, F.: 'Segmentation-Aware Deformable Part Models', *IEEE Conf. Comput. Vision Patt. Recog.*, 2014, pp. 168-175.
9 Li, B., Wu, T. and Zhu, S.: 'Integrating Context and Occlusion for Car Detection by Hierarchical And-Or Model', *Eur. Conf. on Comput. Vision*, 2014, pp. 652-667
10 SanMiguel, J. and Martínez, J.: 'An ontology for event detection and its application in surveillance video', *IEEE Int. Conf. on Advanced Video and Signal-based Surveillance*, 2009, pp. 220-225
11 Wolf, C., Mille, J., Lombardi, E. Celiktutan, O., Jiu, M., Dogan, E., Eren, G., Baccouche, M., Dellandrea, E., Bichot, C., Garcia, C., Sankur, B.: 'Evaluation of video activity localizations integrating quality and quantity measurements', *Comp. Vis. Image Und.*, 2014, **127**, (1), pp. 14-30
12 SanMiguel, J., Escudero-Vinolo, M., Martínez, J., Bescós, J.: 'Real-time single-view video event recognition in controlled environments', *Int. Workshop on Content-Based Multimedia Indexing*, 2011, pp. 91-96.
13 García-Martín, A., Alcedo, B., and Martínez, J.: 'PDbm: people detection benchmark repository', *Electron. Lett.*, 2015, **51**, (7), pp. 59-60