Multi-configurations for part-based person detectors

Alvaro Garcia-Martin[1], Rubén Heras Evangelio[2], Thomas Sikora[2]

Video Processing and Understanding Lab at Universidad Autonoma de Madrid[1], Communication Systems Group at Technische Universität Berlin[2]

People detection is a task that has generated a great interest in the computer vision and specially in the surveillance community. The ability to detect people in video and in particular detecting people in crowded scenarios is the key to a number of multiple applications including video surveillance, group behavior modeling, crowd disaster prevention, etc. Due to the rise in popularity of these applications over the last years, people detection has gradually experienced a great development. Currently, many different systems exist which try to solve the problem posed by the task of detecting people. The state of the art includes several successful solutions working in specific and constrained scenarios. However, the detection of people in real world scenarios such as airports, malls, etc, is still a highly challenging task due to the multiple appearances that different persons may have, heavy occlusions, specially in crowded scenarios, view variations and background variability.

The ongoing work that we want to present and discuss in the "Parts and Attributes" workshop is focused on the improvement of people detection in crowded scenarios. To that aim, we use a part-based person model and propose a statistical driven way of combining the individual body part detectors in order to detect persons even in the case of failure on the detection of any of the body parts. The objective is to be able to detect people with nearly the same reliability whether if they are completely visible (people in front of the group) or only partially visible (people behind).

The most robust approaches regarding the support of different poses and partial occlusions are those based on complex, which define a person as collection of multiple combinable regions or shapes, i.e., part-based models [2, 5]. However, even such approaches have difficulties in dense environments. The Implicit Shape Model (ISM) of [5] is improved in [6] by using a probabilistic formulation in order to generate a model that is scalable from a general object–class detector into a specific object–instance detector, thus making the detection more reliable. The detector in [2] is improved in [3] by using a grammar model which includes an additional "body part" simulating possible occlusions. Also based on [2], in [7] a joint model is proposed, which is trained to detect single people as well as pairs of people under varying degrees of occlusion.

Most closely related to our work is the approach in [3], which demonstrates the advantages of taking into account in the person model the possibility of failure or occlusion of some body parts. In our case, we do not specifically train the model to capture specific occlusion patterns. Instead, we define a more generic scheme in which the absence of any particular body part can be modelled by defining multiple configurations of the part-based models learned during the training phase. Therefore, we are able to deal with occlusions by automatically selecting which of all the possible person model configurations adjust better to any kind of occlusion. In particular, we aim at solving the problem posed to the approach in [3] by crowded scenarios, where the range of possible different occlusions is much bigger and, therefore, the complexity of the grammar model and its training increases exponentially.

The detector in [2] is a part-based person model. It consists of mixtures of multi-scale deformable part models in a star-structure defined by a root model, where the root and each of the deformable body parts are modeled by a HOG as firstly proposed in [1].

The detector proposed in [2] defines N body parts positioned around the root filter ($n = 0$), which models the appearance of the whole body. The N body parts are computed at twice the resolution in relation to the root filter in order to refine the detection based only on the root information. Each of the n detectors, included the root ($n = 0, ..., N$), is modeled by a 3-tuple (F_n, $v_{n,0}$, d_n), where F_n is the HOG filter response (detection confidence) for part n; $v_{n,0}$ is a two-dimensional vector defining the relative position of part n with respect to the anchor position (x_0, y_0) of the root; and d_n is a four-dimensional vector specifying coefficients of a quadratic function defining the cost for each possible placement of the part relative to the anchor position. The $BP_n(x, y, s)$ represents the confidence at pixel position (x, y) for body part n ($n = 0, ..., N$) associated to scale s ($s = 1, ..., S$). Thus, the confidence score for part n at scale s is given as

$$BP_n(x, y, s) = F_n(x, y, s) - \langle d_n, \phi(dx_n, dy_n) \rangle \tag{1}$$

with

$$(dx_n, dy_n) = (x_n, y_n) - (2(x_0, y_0) + v_{n,0}) \tag{2}$$

giving the displacement of part n relative to the anchor and

$$\phi(dx, dy) = \left(dx, dy, dx^2, dy^2\right) \tag{3}$$

defining the potential spatial deformation distributions.

The final detection confidence or score $C(x, y, s)$ is computed as the sum of the root and N body parts at each pixel position and scale.

$$C(x, y, s) = \sum_{n=0}^{N} BP_n(x, y, s) \tag{4}$$

The final multi-scale detection hypotheses are extracted after a thresholding followed by a non-maximum suppression process, used to eliminate possible repeated detections. The chosen threshold or minimum score required in order to consider the detected object as a person depends directly on the total number of body parts detections.

Since the total score depends tightly on the number of parts detected, this approach is not able to reliably cope with occlusions. Therefore, it fails to detect people in groups, where most of the persons are only partially visible.

In order to cope with any kind of body part occlusion, we propose to use multiple person model configurations t ($t = 1, ..., T$) with $1 \leq T \leq 2^N$, where each person model configuration t consist of a subset of M body parts ($m = 1, ..., M$), with $m \subset n$ of the original detector [2] and $1 \leq M \leq N$. Thus, the confidence for each configuration is defined as

$$C_t(x, y, s) = \sum_{n=0}^{N} \alpha_n^t \cdot BP_n(x, y, s) \tag{5}$$

where α^t is a binary selector vector for each configuration t

$$\alpha_n^t = \begin{cases} 1 & , \ n \subset t \\ 0 & , \ otherwise \end{cases} \tag{6}$$

As in the base algorithm, the final multi-scale detection hypotheses are extracted after a thresholding followed by a non-maximum suppression process in order to eliminate possible repeated

detections. However, there are two main differences in our approach with respect to the base algorithm. In first place, there is not only one detection threshold, but there is one for each configuration. Each minimum score required is chosen to be coherent with the number and kind of body parts taken into consideration (see below). In second place, we apply the non-maximum suppression process to the results provided by all the person model configurations togehter.

Once defined the different person model configurations, it is necessary to determine the decision threshold or minimum score required for each configuration in relation to the threshold used if considering the whole set of body parts. To that aim, we consider the confidence or score of each body part n as a continuous random variable BP_n and its associated probability density function $f_{BP_n}(bp_n)$, the final detection confidence as a continuous random variable C and its associated probability density function $f_C(c)$. The minimum confidence k required to consider a detected object as a person corresponds to the probability of $F_C(k) = P(C \leq k)$.

Analogously, each configuration confidence can be considered as a continuous random variable C_t with an associated probability density function $f_{C_t}(c_t)$. In order to estimate the minimum confidence k_t required for each configuration, it is necessary to determine a correction factor R_t that takes into account the number of body parts included in each configuration and their respective contribution or information relevance in relation to the original configuration with N parts. For example, assuming that all the body parts had the same contribution the correction factor $R_t = \frac{1}{N}$ could be used for each configuration t. Nevertheless, since the individual part detectors are not equally discriminative, their contribution to the overall model can not be conisdered the same.

Therefore, in order to estimate the contribution of each body part n, we first estimate the similarity of the distribution of the scores obtained by using the configuration with the whole set of body parts (F_C), with the distribution of the scores obtained by using the configuration with all except the considered body part n. To that aim, we are using at the moment the Kullback-Leibler Divergence (D_{KL}) [4]. We define the similarity KL_nbetween each body part BP_n to the complete model C as the Kullback-Leibler Divergence between the distribution F_C and the distribution without that body part n, $F_{C'}$:

$$KL_n = D_{KL}(F_C||F_{C'}), \text{ being } C' = \sum_{i=0, i\neq n}^{N} BP_i \tag{7}$$

This measure is normalized $K\bar{L}_n$ so that $\sum_{n=1}^{N} K\bar{L}_n = 1$. Finally, the correction factor R_t is computed as the accumulative body parts contributions:

$$R_t = \sum_{n=0}^{N} \alpha_n^t \cdot K\bar{L}_n \tag{8}$$

A factor of $R = 1$ means that there is not necessary any correction on the decision threshold because the considered configuration corresponds to the use of all the body parts.

The minimum confidence k_t required for each configuration t with associated probability $F_{C_t}(k_t) = P(C_t \leq k_t)$ is modified according to the original person model confidence k and the corresponding correction factor R_t:

$$F_{C_t}(k_t) = 1 - R_t(1 - F_C(k)) \tag{9}$$

Therefore, the final probability $F_{C_t}(k_t)$ required for each configuration is defined between the original $F_C(k)$ and 1 ($F_C(k) \leq F_{C_t}(k_t) \leq 1$). The simpler the person model (less body parts), the higher the probability (i.e., the confidence) required to detect a person and vice versa.

In the poster to be presented at the workshop we want to present our very encouraging preliminary results as well as discuss further measures of divergence, which migh help to better correct the probability required for each of the considered configurations.

References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. of CVPR. pp. 886–893 (2005)
2. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9), 1627–1645 (2010)
3. Girshick, R.B., Felzenszwalb, P.F., Mcallester, D.: Object detection with grammar models. In: Proc. of NIPS (2011)
4. Kullback, S., Leibler, R.A.: On information and sufficiency. The Annals of Mathematical Statistics 22(1), 79–86 (03 1951), http://dx.doi.org/10.1214/aoms/1177729694
5. Leibe, B., Seemann, E., Schiele, B.: Pedestrian detection in crowded scenes. In: Proc. of CVPR. pp. 878–885 (2005)
6. Seemann, E., Fritz, M., Schiele, B.: Towards robust pedestrian detection in crowded image sequences. In: Proc. of CVPR. pp. 1–8 (June 2007)
7. Tang, S., Andriluka, M., Schiele, B.: Detection and tracking of occluded people. International Journal of Computer Vision (2014)