Electronics Letters 49.4 (2013): 255-256

DOI: http://dx.doi.org/10.1049/el.2012.3817

Enhanced people detection combining appearance and motion information

Journal:	*Electronics Letters*
Manuscript ID:	ELL-2012-3817
Manuscript Type:	Letter
Date Submitted by the Author:	30-Oct-2012
Complete List of Authors:	Garcia-Martin, Alvaro; Universidad Autonoma Madrid,Escuela Politecnica Superior, Departamento de Tecnología Electrónica y de las Comunicaciones Martinez, Jose M.; Universidad Autonoma Madrid,Escuela Politecnica Superior, Departamento de Tecnología Electrónica y de las Comunicaciones
Keywords:	People detection, Discriminatively Trained Deformable Parts, Implicit Motion Model, MoSIFT

Enhanced people detection combining appearance and motion information

A. García-Martín and J.M. Martínez

The combination of two of the most recent people detectors from the state of the art is proposed. It is already known that the combination of independent information sources is useful for any detection task. In relation with people detection, there are two main discriminative information sources that characterize a person: appearance and motion. We propose the combination of two recent approaches based on both information sources. Experimental results over an extensive dataset show that the proposed combination significantly improves the results.

Introduction: In video analysis, people detection consists of locating all persons present in a scene according to a previously defined person model. People detection is one of the most challenging problems in computer vision. The complexity of the people detection problem is mainly based on the difficulty of modeling persons because of their huge variability in physical appearances, articulated body parts, poses, movements, points of views and interactions between different people and objects. This complexity is even higher in real world scenarios such as airports, malls, etc, which often include multiple persons, multiple occlusions and background variability.

As already presented in [1,2], our framework for people detection is based on the combination of appearance and motion information in order to work in more complex or realistic scenarios. In particular, in this letter we introduce in our framework the use of one of the most successful appearance-based approaches for general object detection to date [3] in combination with our people detector based on motion. The experimental results show how the improvement in one of the information sources affects positively to the final combination results.

People detection: The proposed people detection framework was already described in [1] and extended in [2]. It is able to perform two independent visual people detections, the first one using the shape or appearance of humans as discriminative feature and the second one using their motion. The final detection result is the combination of both sources outputs. The appearance and motion detections have been combined at blob level (position and dimension) following the Multiple Hypotheses Simplification Criteria (MHSC) [2].

People detector based on appearance: The appearance people detector is based on the Discriminatively Trained Deformable Parts (DTDP) [3]. DTDP is a part-based extension of the traditional Histogram of Oriented Gradients (HOG) detector [4]. Let us consider the part-based multi-scale detector (Fig. 1), where $P_n(x,y,s)$ represents the confidence at pixel position (x,y) for body part n $(n=1,\dots,N)$ associated to scale s $(s=1,\dots,S)$. Let also each body part be modelated by a 3-tuple (F_n, v_n, d_n), where F_n is the HOG filter response [4] for part n; v_n is a two-dimensional vector defining the relative position of part n with respect to the anchor position (x_0, y_0) of the whole body; 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 confidence score for part n at scale s is given as

$$P_{n(x,y,s)} = F_n(x,y,s) - \langle d_n, \Phi(dx_n, dy_n)\rangle \quad (1)$$

with

$$(dx_n, dy_n) = (x_n, y_n) - (2(x_0, y_0) + v_n) \quad (2)$$

giving the displacement of part n relative to the anchor and

$$\Phi(dx, dy) = (dx, dy.dx^2, dy^2) \quad (3)$$

defining the potential spatial deformation distributions [3].

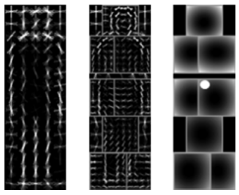

Fig. 1 *Multi-part person model and anchor position* (x_0, y_0) *from [3].*

People detector based on motion: The motion people detector is based on the Implicit Motion Model (IMM) [1]. IMM is a variation of the Implicit Shape Model (ISM) [5], the main difference consist of the use of motion features (MoSIFT) instead of shape features (SIFT). The IMM consists of a codebook C_{IMM} of local motions, that are prototypical for the object category and a spatial probability distribution $P^{C_{IMM}}$ which specifies where each codebook entry may be found on the object. The K elements of C_{IMM} are the motion part of MoSIFT descriptors $d_1^{C_{IMM}}, \dots, d_K^{C_{IMM}}$ extracted around scale-invariant and spatio-temporal interest points (x_k^t, y_k^t, s_k^t). The codebook C_{IMM} is generated using the Reciprocal Nearest Neighbors (RNN) clustering algorithm and the spatial probability distribution $P^{C_{IMM}}$ is learned using annotated training sequences or pairs of images.

Given a new test pair of images, the new features are matched to the learned codebook C_{IMM} in multiple clusters with different weights. Each matching casts votes for theoretical positions of the person centre according to the corresponding learned spatial distribution $P^{C_{IMM}}$. Then, the hypotheses are defined as local maxima in the voting space (x,y,s). Assuming symmetry with respect to our hypothetical centres, a bounding box is obtained for each hypothesis. Finally, multiple hypotheses with more than 50% cover and overlap, as defined in [5], are simplified to the highest score one.

Experimental results: In order to evaluate the performance of the proposed approach (DTDP+IMM) and compare the results with our previous work [2] (ISM+IMM, HOG+IMM, TUD+IMM), we evaluate the new appearance and motion combination following the same methodology, i.e., the same evaluation dataset and evaluation metrics [6].

The test dataset contains 36 surveillance annotated sequences (3698 frames). This dataset contains highly crowded scenes, severely cluttered background and people at different scales.

As in our previous work [1,2], focusing on the people detection evaluation in video surveillance systems, firstly we have evaluated (Precision, Recall and F1Score) each separate detector and then their fusion over the 36 test sequences. Table 1 summarizes the obtained detection results. Using only the appearance information, the DTDP detector gets the best results in terms of Precision (95.1%), Recall (19.5%) and F1Score (30.7%). As in all the other cases, the fusion of two independent information sources (DTDP+IMM) provides a significant improvement in terms of Recall (34.9 or 117.4%) and F1Score (27.7 or 84.9%), but also as in the case of the HOG detector, there is a slightly improvement in terms of Precision (1.4%). Finally, using the combination of appearance and motion information, the detector DTDP+IMM gets the best results in terms of Precision (96.4%), Recall (26.3%) and F1Score (39.2%), mainly due to the better results of the DTDP detector by itself.

Figs. 2*a*, *b* and *c* show people detection examples of both approaches (DTDP and IMM) and also the final combination result (DTDP+IMM). It is clear that appearance information is much more efficient than the motion information mainly in terms of Recall (19.5% vs. 12.1%), but even so the motion information is useful in cases where appearance is not enough.

Table 1: Detection results. Percentage increase (%Δ) calculated with respect to single appearance versions (Appearance+Motion) or with respect to the single motion version (Motion+Appearance).

	Approach	Precision (%)	%Δ	Recall (%)	%Δ	F1Score (%)	%Δ
Appearance	ISM	94.7	-	16.5	-	27.2	-
	HOG	93.4	-	15.4	-	25.3	-
	TUD	95.1	-	10.7	-	19.1	-
	DTDP	95.1	-	19.5	-	30.7	-
Motion	IMM	95.1	-	12.1	-	21.2	-
Appearance + Motion	ISM+IMM	93.9	-0.8	21.7	+31.5	34.6	+27.2
	HOG+IMM	96.4	+3.2	21.5	+39.6	34.5	+36.4
	TUD+IMM	94.4	-0.7	17.0	+58.9	28.5	+49.2
	DTDP+IMM	**96.4**	**+1.4**	**26.3**	**+34.9**	**39.2**	**+27.7**
Motion + Appearance	ISM+IMM	-	-1.3	-	+79.3	-	+63.2
	HOG+IMM	-	+1.4	-	+77.7	-	+62.7
	TUD+IMM	-	-0.7	-	+40.5	-	+34.4
	DTDP+IMM	-	**+1.4**	-	**+117.4**	-	**+84.9**

a

b

c

Fig. 2 *People detection example results. (a) DTDP results, (b) IMM results and (c) DTDP+IMM results.*

Conclusion: It has been already demonstrated that the combination of appearance and motion information is useful for people detection in complex scenarios. However, we propose a new appearance and motion combination using two recent people detection approaches from the state of the art. Experimental results show that the DTDP detector achieves better detection results than the other approaches from the state of the art based only on appearance. Consequently, our new proposed combination also achieves better final detection results.

Acknowledgments: This work was partially supported by the Universidad Autónoma de Madrid ("FPI-UAM") and by the Spanish Goverment ("TEC2011-25995 EventVideo").

A. García-Martín and J.M. Martínez (Video Processing and Understanding Lab, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Ciudad Universitaria de Cantoblanco, 28049 Madrid, Spain)
E-mail: {alvaro.garcia, josem.martinez}@uam.es

References:

1 A. Garcia-Martin, A. Hauptmann and J.M. Martinez. People detection based on appearance and motion models. In Proc. of AVSS, 2011, pp. 256-260.

2 A. Garcia-Martin and J. M. Martinez. On collaborative people detection and tracking in complex scenarios. Image and Vision Computing, May 2012, Vol. 30 (4-5), pp. 345-354.

3 P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, September 2010, Vol. 32(9), pp. 1627-1645.

4 N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In Proc. of CVPR, 2005, pp. 886-893.

5 B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In Proc. of CVPR, 2005, pp. 878-885.

6 A. Garcia-Martin, J.M. Martinez and J. Bescos. A corpus for benchmarking of people detection algorithms. Pattern Recognition Letters, January 2012, Vol. 33 (2), pp. 152-156.