People Detection Based on Appearance and Motion Models

Álvaro García-Martín
Video Processing and Understanding Lab
Universidad Autónoma de Madrid (Spain)
alvaro.garcia@uam.es

Alex Hauptmann
School of Computer Science
Carnegie Mellon University (USA)
alex@cs.cmu.edu

José M. Martínez
Video Processing and Understanding Lab
Universidad Autónoma de Madrid (Spain)
josem.martinez@uam.es

Abstract

The main contribution of this paper is a new people detection algorithm based on motion information. The algorithm builds a people motion model based on the Implicit Shape Model (ISM) Framework and the MoSIFT descriptor. We also propose a detection system that integrates appearance, motion and tracking information. Experimental results over sequences extracted from the TRECVID dataset show that our new people motion detector produces results comparable to the state of the art and that the proposed multimodal fusion system improves the obtained results combining the three information sources.

Keywords: People detection, implicit shape model, implicit motion model, MoSIFT

1. Introduction

In recent years, computer vision has seen great progress. It is an evolving field with multiple lines of research and application. People detection is one of the most challenging problems in this field. 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. At the same time, people detection has a wide range of applications including video surveillance, intelligent systems (robotic), image and video indexing, driver assistance systems, video games, etc.

Currently, many different systems exist which try to solve this problem. The state of the art in people detection and tracking includes several successful solutions working in specific and constrained scenarios. Over the last few years, there have been multiple approaches in more realistic environments with multiple people and occlusions [2, 12], and even onboard scenarios [17]. Most of them get acceptable results using only the appearance information or adding tracking information. To achieve a more reliable performance, we propose to combine people detection based on appearance, people detection based on motion and tracking information.

The main contribution presented in this paper, is a new motion model inspired by the well-established ISM people detection approach [11] and the MoSIFT descriptor [4], successfully employed in activity recognition. Combining both ideas, a new people detection approach based on their motion is introduced: Implicit Motion Model (IMM). Furthermore, to evaluate this new detector, a full system that combines appearance, motion and tracking information has been designed/developed.

The remainder of this paper is structured as follows: section 2 presents a brief of the state of the art, section 3 overviews our complete system which combines different information sources, whilst section 4 describes our new people motion model. Section 5 describes the experimental dataset and results. In section 6, the main conclusions are summarized and future work is described.

2. State of the art

In the following, we give an overview of current people detection approaches, focusing on the kind of information they employ: appearance and/or motion. There is a more comprehensive study of the use of appearance information in the state of the art, mainly due to the fact that appearance

provides us a much more discriminant information about people detection. There are some approaches that include motion information to add robustness to the detection and there are very few cases where the only information used is motion.

Most of the existing approaches are only based on appearance information. There are two major types of approaches based on appearance: On the one hand, the methods based on simplified person models (only a region or shape) [6, 11, 19, 20]. [6] uses a person model based on HOG (histograms of oriented gradients) descriptors and an SVM classifier, [11] makes use of shape representation with the generative ISM framework, [19] uses an ellipse model and a silhouette fitting algorithm and [20] performs the classification by similarity with silhouettes stored in a codebook. On the other hand, there are methods based on combination of multiple parts [3, 9, 18, 8]. [3] trains multiple detectors for anatomically defined body parts which are then combined using pictorial structures, [9] performs an analysis of concavity and convexity of the silhouette to identify different body parts, [18] tries to identify the characteristic edges of a human body and to generate four edge models (body, head, torso and legs), each model is trained using a nested Adaboost cascade structure and [8] proposes a real-time adaptation of the work presented in [18].

It is known that human motion is an important cue for people detection. However, there are not many approaches that make use of this information. Some authors combine appearance and motion expanding their own previous works to more than one frame [7, 16]; they improve significantly the results, but do not generate a motion model as an independent entity. Some approaches use only the motion information [5, 15]. [5] applies time-frequency analysis to detect and characterize the human periodic motion and [15] detects patterns of human motion using optical flow and an SVM classifier.

3. System overview

A complete framework has been designed to predict/update the visual people detection, see Figure 1. 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. Using the people detection as first step, the framework is able to update the person detection (appearance, motion or their fusion) iteratively over time using a color based tracker.

Appearance model

The appearance people detector is based on the ISM [11]. An ISM is a generative model for object detection and has been applied to a variety of object categories, including cars, motorbikes, animals and pedestrians. It consists of a codebook C of local appearances that are prototypical

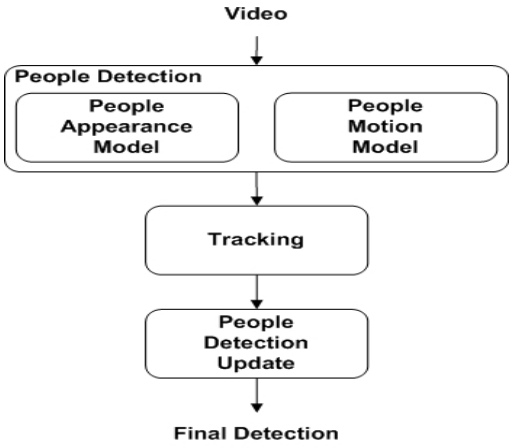

Figure 1. System Overview

for the object category, and a spatial probability distribution P_C which specifies where each codebook entry may be found on the object. The K elements of C are local descriptors $d_1^C .....d_K^C$ extracted around scale-invariant interest points (x_k, y_k, s_k), the codebook C is generated using an agglomerative clustering with average linkage and only the cluster centers are stored. The spatial probability distribution P_C is learned during a second training phase where all the local descriptors are matched in multiples clusters with different weights.

Motion model

Our motion model is based on the ideas of the ISM framework. The substantial difference is the use of motion rather than shape information. Details of our algorithm are described in the following section.

Tracking

In our system, tracking is based on [14], adding to the particle filter algorithm an adaptive appearance model based on color distributions. The object model is represented by a weighted histogram which takes into account both the color and the shape of the target. It also includes a straightforward kinematic system model to propagate the particle filter sample set. The observation probability of the particle filter mean state will be used as tracker confidence level C_t in the people detection update.

People detection update

A tracking process is initialized for each detected person. The following detections will update existing trackers or will create new tracking processes. The conditional probability of people detection, given the tracking information in each frame $P_{p|t}(f)$, will be predicted/updated over time based on current people detection probability $P_p(f)$ and the tracker confidence level $C_t(f)$:

Figure 2. SIFT (left) and MoSIFT (right) interest points. Yellow circles indicate interest points and their scales, red arrows indicate the dominant motion orientation.

Figure 3. Detection process examples. Voting space (black lines), center hypotheses (green points), hypotheses (red rectangles) and final hypothesis (green rectangles)

$$P_{p|t}(f) = \begin{cases} P_p(f), & P_p(f) > 0 \\ P_{p/t}(f-1) - (1 - C_t(f)), & P_p(f) = 0 \end{cases} \quad (1)$$

4. People motion model

The pattern of human motion is well known to be readily discriminative from other types of motions [5, 7, 16]. We introduce a new human motion representation that is mainly based on the use of the ISM framework [11] and the motion information in the MoSIFT descriptor [4].

4.1. MoSIFT

MoSIFT [4] is a variation of the well-known SIFT point detector and descriptor [13]. MoSIFT detects interest points and encodes not only their local appearance but also explicitly local motion. It consists in tree main steps: firstly, the SIFT algorithm is applied to find scale-invariant interest points in the spatial domain, then optical flow is extracted around the distinctive points with (temporal) motion constrains at corresponding scales and finally the feature descriptor is generated. Figure 2 shows results of SIFT and MoSIFT over the same frame.

In order to generate the feature descriptor, MoSIFT adapts the idea of grid aggregation in SIFT to describe motion, but instead of using appearance gradients, it uses the optical flow. The other main difference to appearance description is in the rotation invariance. Rotation invariance is important to appearance since it provides a standard to measure the similarity of two interest points, but the direction of movement is actually an important (non-invariant) vector to discriminate different movements. The two aggregated histograms (appearance and optical flow) are combined into the MoSIFT descriptor, which has therefore 256 (128+128) dimensions.

4.2. Implicit Motion Model

The main idea consists of identifying and learning characteristic motions of humans in typical surveillance systems and generating a motion model. We propose to use the motion information in the MoSIFT descriptor to characterize the movements and build a motion model following the ISM framework.

4.2.1 Learning the Motion Model

For symmetry with the ISM model, the Implicit Motion Model $IMM(C) = (C, P_C)$ consists of a codebook C of motion appearances that are prototypical for the object category, and a spatial probability distribution P_C which specifies where each codebook entry may be found on the object. The K elements of C are the motion part of MoSIFT descriptors $d_1^C .....d_K^C$ extracted around scale-invariant and spatio-temporal interest points (x_k^t, y_k^t, s_k^t), the codebook C is generated using RNN (reciprocal nearest neighbor) clustering algorithm [11] and the spatial probability distribution P_C is learned using annotated training sequences or pairs of images, our training dataset includes several sequences but other datasets only include pairs of images which are enough for training the motion model.

4.2.2 Detection process

Given a new test pair of images, the SIFT interest point detector is applied again and MoSIFT features are extracted around the selected locations. Then these features are matched to the learned codebook C in multiple clusters with different weights. Each matching casts votes for theoretical positions of the person center according to the learned spatial distribution P_C. Then, the hypotheses are defined as local maxima in the voting space (x, y, s). Assuming symmetry with respect to our hypothetical centers, a bounding box (blob) is obtained for each hypothesis. Finally, multiple hypotheses with more than 50% cover and overlap, as defined in [12], are simplified to the highest score one. Figure 3 shows two examples of the same sequence.

5. Experimental results

This section describes the experimental dataset (training and test dataset) and the results obtained in each stage of

Figure 4. Ground truth dataset examples

our system.

5.1. Experimental dataset

In order to evaluate the performance of the proposed approach, we introduce a video dataset containing 61 surveillance annotated sequences (6353 frames). These sequences have been extracted from TRECVID 2008 dataset [1]: the sequences for the surveillance event detection task were recorded at London Gatwick International Airport. This dataset contains highly crowded scenes, severely cluttered background and people at different scales. The references, descriptions and annotations of the sequences are freely available for academic purposes (**http://www-vpu.ii.uam.es/PDds/**).

Due to the small size of the objects at the top of the image, during the annotation of sequences and the evaluation of algorithms only the 85% of the bottom of the images has been taken into account.

Training dataset

Our IMM has been trained with 25 sequences (2655 frames). Each sequence includes multiple annotated people but the IMM has been trained using only the person with maximum motion information (MMI) per video. The MMI person has a trajectory completely non-occluded since entering the scene until they come out of it. The MMI person has been manually selected in each video.

Test dataset

The test set is composed of 36 sequences (3698 frames). All people present at the scene have been manually annotated and have been taken into account in the evaluation. Figure 4 shows some examples of final annotations.

5.2. People detection results

In order to evaluate the different people detectors and integrated system, firstly we have evaluated each separate detector and their fusion over the 36 test sequences. The appearance and motion detectors have been combined at blob level: both detectors have been run independently and the results (blobs) have been added, or have been averaged in those cases of overlapping blobs. The ISM results have been obtained using author's code and binaries (**http://www.vision.ee.ethz.ch/~bleibe/index.html**) and the IMM has been implemented using the LIBPMK library [10] .

	Precision	Δ	Recall	Δ	F1Score	Δ
ISM	94.7	0.0	16.5	0.0	27.2	0.0
IMM	95.3	+0.6	12.1	-26.7	21.2	-22.1
ISM+IMM	93.9	-0.8	21.7	+31.5	34.6	+27.2

Table 1. Detection results. Percentage increase (Δ) calculated on ISM.

We can see in Table 1 the average results for the test data. We can see how both algorithms with high precision values (~94%) differ from recall values (12~16%). It is logical that the motion-based detector obtains lower recall values because only moving people can be detected. However, in environments as complex as these ones, the use of motion information obtains results close to the use of appearance information. The combination of both detectors obtains better recall results (21.7%), slightly reducing precision values (93.9%).

Secondly, we have evaluated the whole system over the same 36 test sequences. Using algorithms with high precision values (~94%), our prediction/update based on tracking confidence is able to maintain high precision values (91.8~93.7%) but improving considerably the recall (18.6~28.4%). This process exploits the intermittent operation of the detectors. The false positive detections with usually lower people detection confidence will be discarded in a few frames. However, the intermittent true positives detections will be expanded over time using tracking predictions. We can see in Table 2 the average results of three different system configurations, the ISM detector, the IMM detector and their fusion, all of them adding the tracking information.

Every video surveillance system and/or people detector must maintain a compromise between precision and recall. Thinking about the people detection as a preliminary step in the event detection task (e.g., TRECVID Surveillance event detection), it is more valuable to get better recall results at the expense of getting slightly reduced precision results. At higher semantic levels (activity recognition/detection), the people detection false positives can be easily dismissed, but on the other hand the undetected people can not be recovered.

	Precision	Δ	Recall	Δ	F1Score	Δ
ISM+Tracking	93.7	-1.1	22.8	+38.2	37.4	+37.5
IMM+Tracking	93.1	-2.3	18.6	+53.7	32.1	+51.4
ISM+IMM+ Tracking	91.8	-2.2	28.4	+30.9	44.6	+28.9

Table 2. System results. Percentage increase (Δ) calculated on each approach without tracking.

6. Conclusions

In this paper, a new people detection motion model IMM is proposed. Using the ISM Framework and the MoSIFT interest points detector and descriptor, we present a new people detection algorithm based in characteristic movements of people. It is clear that human motion provides useful information for people detection and independent from appearance information, so we also present an integrated system which combines an appearance people model, our new motion model and a tracking algorithm. Experiments have been conducted on challenging sequences extracted from the TRECVID dataset. The results show that our motion-based detector produces results comparable to the ISM state of the art approach. The evaluation of the whole system shows how the combination of different information sources improves the final detection, obtaining a significant improvement in recall and a slightly precision reduction.

In the future, we propose the study of different fusion/combination techniques between the appearance and motion detectors to improve the recall without compromising the precision, or even the creation of a single integrated Implicit Shape-Motion Model (ISMM), using the full MoSIFT description.

7. Acknowledgments

This work has been partially supported by the Cátedra UAM-Infoglobal ("Nuevas tecnologías de vídeo aplicadas a sistemas de video-seguridad") and by the Universidad Autónoma de Madrid ("FPI-UAM: Programa propio de ayudas para la Formación de Personal Investigador")

References

[1] Trecvid 2008 evaluation for surveillance event detection. In *http://www.nist.gov/speech/tests/trecvid/2008/doc/EventDet08-EvalPlan-v04.htm*. National Institute of Standards and Technology (NIST), 2008. 4

[2] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *Proc. of CVPR*, pages 1–8, 2008. 1

[3] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Proc. of CVPR*, pages 1014–1021, 2009. 2

[4] M.-Y. Chen and A. Hauptmann. Mosift: Recognizing human actions in surveillance videos. Technical Report CMU-CS-09-161, Carnegie Mellon University, 2009. 1, 3

[5] R. Cutler and L. S. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8):781–796, 2000. 2, 3

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. of CVPR*, volume 1, pages 886–893, 2005. 2

[7] N. Dalal and B. Triggs. Human detection using oriented histograms of flow and appearance. In *Proc. of ECCV*, 2006. 2, 3

[8] A. Garcia-Martin and J. Martinez. Robust real time moving people detection in surveillance scenarios. In *Proc. of AVSS*, pages 241–247, 2010. 2

[9] I. Haritaoglu, D. Harwood, and L. S. Davis. Ghost: a human body part labeling system using silhouettes. In *Proc. of ICPR*, volume 1, pages 77–82, 1998. 2

[10] J. J. Lee. Libpmk: A pyramid match toolkit. Technical Report MIT-CSAIL-TR-2008-17, MIT Computer Science and Artificial Intelligence Laboratory, April 2008. 4

[11] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision*, 77:259–289, 2008. 1, 2, 3

[12] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *Proc. of CVPR*, pages 878–885, 2005. 1, 3

[13] D. Lowe. Distinctive image features from scale invariant key points. *International Journal of Computer Vision*, 2004. 3

[14] K. Nummiaro, E. Koller-Meier, and L. V. Gool. An adaptive color-based particle filter. *Image and Vision Computing*, 21:99–110, 2003. 2

[15] H. Sidenbladh. Detecting human motion with support vector machines. In *Proc. of ICPR*, pages 188–191, 2004. 2

[16] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *Proc. of ICCV*, volume 2, pages 734–741, 2003. 2, 3

[17] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *Proc. of CVPR*, 2009. 1

[18] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *Proc. of ICCV*, pages 90–97, 2005. 2

[19] F. Xu and K. Fujimura. Human detection using depth and gray images. In *Proc. of AVSS*, pages 115–121, 2003. 2

[20] J. Zhou and J. Hoang. Real time robust human detection and tracking system. In *Proc. of CVPR*, page 149, 2005. 2