IMPROVING MULTI-CAMERA PEOPLE DETECTION USING CONTEXTUAL INFORMATION

Rafael Martín-Nieto, Alejandro Miguélez-Sierra, Alvaro García-Martín, José M. Martínez

Video Processing and Understanding Lab (VPULab), Universidad Autónoma de Madrid, Spain

ABSTRACT

This paper focuses on the improvement of people detection algorithms by using contextual information in order to combine information obtained from different cameras. Using multiple cameras and information from the recorded scenario, called contextual information (distances between detected objects and cameras, position of the cameras, etc.), it is possible to improve the detection performance taking advantage of the results of the other cameras, transferring information from one camera to another, and then combining it. The proposed system has been evaluated on two different datasets obtaining improvements in the performance of the detection algorithm, both for detections with different appearances and aspect ratios, as well as for a system with different camera orientations.

Index Terms— Computer vision, contextual information, people detection, Multi-camera environment.

1. INTRODUCTION

People detection in video sequences is one of the most complex tasks in the field of video signal processing. The task complexity is mainly due to the difficulty of defining a generic people model, due to the wide variety of appearance and pose. There are multiple detection algorithms in the state of the art, but all them have errors when detecting in certain scenarios with one or multiple factors like: occlusions, illumination changes, perspective changes, etc.

People detection techniques can be divided into three stages [1]:

- First, a person model is designed which defines the characteristics that the detected objects must fulfill to be considered people.
- Second, an object extraction process is performed, which will find the candidates to be classified.
- Finally, the classification consists of the comparison of the objects detected in the sequence with the model generated in the first step. In this step, a decision is made on the objects and it is decided whether the objects are classified as persons or not. Depending on the application, the decision can be person vs. non person, or a probability value of being a person.

The information provided by a single camera is limited, so in order to monitor a wide area or to obtain more information from the different viewpoints of a region of interest, it is necessary to use more than one camera. For this reason, the use of several cameras is a common way of developing applications, since it is also useful for solving occlusions in scenarios with high density of people and for 3D applications.

This work has been partially supported by the Spanish government under the project TEC2014-53176-R (HAVideo) and by the Spanish Government FPU grant programme (Ministerio de Educación, Cultura y Deporte).

The key technologies when using multi-camera are, as explained in [2]:

- Calibration of all cameras in a single coordinated system.
- Knowledge of the topology of the camera network to get information on how they are related to each other.
- Identification of objects in several cameras to determine if the observed objects are the same or different.
- Tracking objects across all cameras.
- Automatic recognition of abnormal actions or activities.

The use of a multi-camera environment in scenarios with possible occlusions, usually improves the detection performance with respect to the use of the cameras independently. A method is proposed in [3] to perform detection and tracking of people in multi-camera environments where there are occlusions. This method is based on the methodology proposed in [4] and is based on using the information of each of the cameras from the scenario, merging it into a common plane (the ground plane) obtained by homographies. The individual information that is combined in the common plane is previously obtained by subtracting the background. Then, objects detection is performed in the common plane and, afterwards, a correspondence is made for all the views of the different cameras to identify the objects in all of them. In this way, using cameras with different locatons, the problem of occlusion is generally solved. The main drawback is that the individuals had to appear initially isolated. In [5], an improvement of the previous method [4] to eliminate false positives is proposed. The algorithm that performs this process compares the views of all the cameras for each one of the detected objects, thus reducing the false detections. It is also interesting to consider [6], where a method that uses a Kalman filter to obtain 3D information from the 2D information is presented.

Unlike the previous approaches, we transfer the detections from one camera to another instead of project all the detections to the common plane, as in this way the object information is not reduced to a simple coordinate, allowing to transfer more information. Instead of processing the information in the common plane, the detections in each camera are combined with the information from the other ones, which also can be used to further combine the information in the common plane (with the advantage of having previously improved the information of each camera), as the traditional approaches of the state of the art.

In the work presented in this paper, the common plane will be used to obtain the different camera views information, and it is used to transfer (and combine) object detections from one camera to another. The main contribution is the consideration of a cylinder to approximate the person position in order to transfer the position of the detection bounding boxes from one camera to another, maintaining the volume that occupies a person, instead of using only the

projected plane generated from the detected bounding box. The consideration of the representation of people as cylinders has been used previously in the state of the art [7], but as a method for people counting (estimation) from a single perspective.

The structure of the following sections is as follows. Section 2 describes the proposed technique for the information transference and combination between cameras. Section 3 presents the evaluation framework. Experimental results are discussed in Section 4. Finally, Section 5 summarizes conclusions and future work.

2. PROPOSED TECHNIQUE

2.1. Cylinder estimation and information transference

The objective of the developed technique is to use the bounding boxes of the detections from one camera and transfer them to the point of view of another camera. As the projections on the common plane of the detected bounding boxes do not correspond spatially with the position of the detected object, the transference between camerasmust be corrected. Figure 1(c) shows a case in which the bounding box transferred (blue) does not fit the person when changing the point of view, so it is necessary to process it to obtain the red box. Here we describe the method applied to each bounding box detected by the camera whose information is transferred.

1. firstly, the base (bottom) segment of the detection bounding box is projected to the common plane. This plane can be obtained using homographic techniques, or from the intrinsic and extrinsic parameters of the cameras. The base segment is used because it is in the common plane, which allows to transfer it in a more precise way. Figure 1(a) shows two bounding boxes which will be transferred.
2. Using the projected segment in the common plane, a circumference is defined so that the projected segment forms one of the sides of a square inscribed therein. In Figure 1(b), the projected segment is represented with the blue line, the square is represented with discontinuous blue line, and the circumference is represented with a (green) circle.
3. To define the bounding box base segment which will be transferred to the other camera, the inscribed square (blue) is rotated (represented with discontinuous red line in Figure 1(b)) with an angle such that the closest side is perpendicular to the line connecting the new camera with the center of the circumference (green cross in Figure 1(b)). This side (red line) corresponds to the projection of the transferred bounding box base segment.
4. The height of the cylinder is estimated using proportionality, taking into account the object original heigth and the cameras distances to the object.
5. Finally, this generated cylinder is transferred to the point of view of the new camera, again using an homography (inverse matrix) or from the intrinsic and extrinsic parameters of the new camera. An example of the resulting cylinders is shown in Figure 1(c).

2.2. Detections combination

It is common for an object to be detected in several cameras, so it is necessary to add an information combination stage. The multiple matching bounding boxes of the same person are simplified into a single one. The measures used for this association are the same as the measures used in the evaluation to decide if two bounding boxes correspond to the same object (see subsection 3.3). Two detection bounding boxes are considered to correspond with the same person if $rd \leq 0.5$ (relative distance between bounding boxes, corresponding to a deviation up to 25% of the true object size) and cover and overlap between bounding boxes are both above 50%. More details of these commonly used metrics are available in [8]. When two bounding boxes are associated as belonging to the same object, the one with greater confidence is maintained, and the other one is deleted to eliminate redundancy.

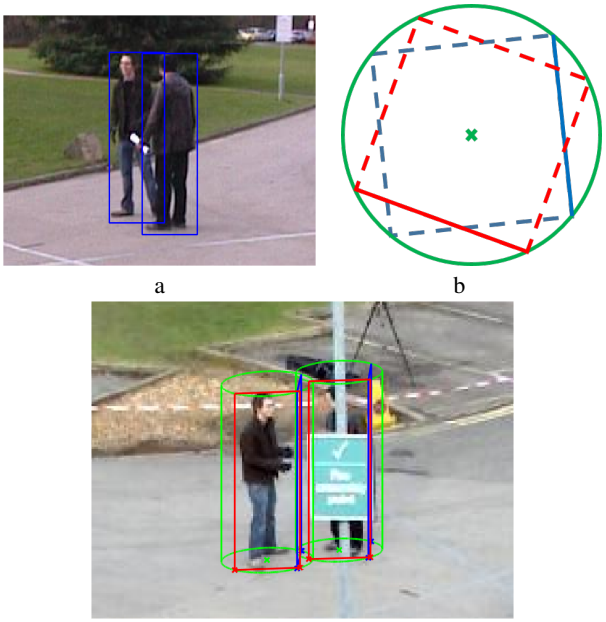

Fig. 1. Overview of the proposed technique process. (a) shows two detection bounding box examples, (b) schematizes the cenital geometric process, and (c) contains a representation of the resulting cylinders (green), the original bounding box (blue, very tilted due to the angle between the cameras viewpoints), and the resulting bounding boxes (red).

Figure 2(c) shows examples of bounding boxes of the ground truth (red), the own camera detections (green) and the transferred detections by the other camera (blue). In this case, each camera separately is capable of detecting only two people (2(a) and 2(b)). For the person in the middle of the image, the bounding boxes of the two cameras (blue and green) are combined into a single one, resulting in a final complete detection containing three bounding boxes, one for each person, each one being very similar to those annotated manually in the ground truth (red). A complete result is obtained, eliminating redundancy between cameras detections.

3. EVALUATION FRAMEWORK

3.1. Detection Algorithm

The method proposed in this paper works at bounding box level, so it can be applied to any detector whose outcome is composed of a list bounding boxes.

Fig. 2. Detection combination example. (a) shows the own camera detections, (b) shows the transferred detections, and (c) represents the own (green), the transferred (blue) and the ground truth (red).

For the evaluation of the method, we have decided to use the algorithm known as DPM ("Deformable Part Models") [9]. The reason to choose this algorithm is that we had a previously trained wheelchair user model, which allowed to evaluate the results on a dataset with different people appearances (WUds, which contains standing people and wheelchair users, see subsection 3.2.1). Given an object, it may be divided into parts having common properties, while the deformable component may be characterized by the connection between pairs of neighboring parts. In case the needed models were available for another detection algorithm, it could be used in the same way as the selected one (e.g. R-CNN [10], Fast R-CNN [11] or Faster R-CNN [12] detectors).

3.2. Datasets

3.2.1. Wheelchair Users dataset (WUds)

This dataset was recorded by the Video Processing and Understanding Lab due to the lack of public wheelchair datasets. The sequences were recorded in a real environment of a senior residence, in order to work with an environment as realistic as possible (due to privacy issues, real recording with actual residents was not possible). Each frame has a resolution of 768x432 pixels and the sequences are recorded at 25 fps. All sequences were recorded in the same room, using two GoPro cameras HERO3 White edition. The fisheye effect was corrected using the GoPro Studio software tool.

The dataset consists of 11 sequences, each of them recorded from two points of views, resulting in a total of 22 sequences. The dataset is composed of 14864 frames with between 1 and 4 people, and up to 4 wheelchair users can appear simultaneously. This dataset and its annotated ground truth are publicly available (`http://www-vpu.eps.uam.es/DS/WUds/`).

Dataset	#Sequences	#Cameras	#Frames	People density
WUds	22	2	214.864 (x2)	From 1 to 4
PETS2009*	2	5	902 (x5)	From 1 to 7

Table 1. Numerical data of WUds andPETS2009 (*subset of sequences with availability of the 5 selected views).

Fig. 3. Frame examples of the Wheelchair Users dataset. (a): viewpoint 1. (b): viewpoint 2.

3.2.2. PETS2009

In order to further evaluate in a more realistic scenario, PETS 2009 Benchmark sequences have been used (`http://www.cvg.reading.ac.uk/PETS2009/a.html`). They are recorded outdoor sequences from a typical surveillance setup.

We evaluate the proposed method (on public available video sequences) with the available ground truth from [13]. The evaluations are done for the View 1 which is the one that has ground truth available, and over region R1 (see website for details), defined by the dataset creators for the use of multiple views. In addition to view 1, views 5, 6, 7 and 8 are also used. The cameras locations are shown in a satellite map in the dataset webpage. We use view 8 as it is the most similar to the scenario of the Wheelchair Users dataset, in which the two cameras used are facing each other. Views 5 and 7 are considered to see the behavior of the proposed technique when the views are (almost) orthogonal. Finally, view 6 is included to check the effect of combining two cameras with the same orientation but at a different distance from the monitored area. Therefore, the sequences that have these five points of view are used, which are S2.L1 and S3.MF.

Table 1 shows numerical data of the datasets.

3.3. Performance measure

Global sequence performance is usually measured in terms of Precision-Recall (PR) curves [14]. These curves compare the similarities between the output and ground truth bounding boxes.The different precision-recall curve values are obtained according to the following formulas:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \quad (1)$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \quad (2)$$

where TruePositives, FalsePositives and FalseNegatives are, respectively, true, false and missed detections.

Additionaly, in order to evaluate not only the yes/no detection decision but also the precise people locations, we take into account the three evaluation criteria defined in [8], that allow to compare hypotheses at different scales: relative distance (rd), cover and overlap.

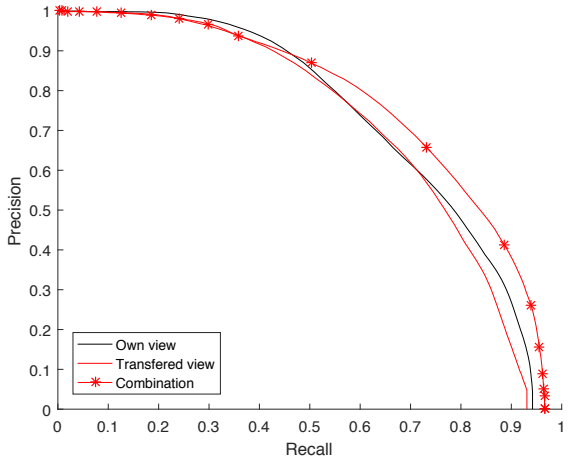

Fig. 4. AUC curves for theWUds evaluation.

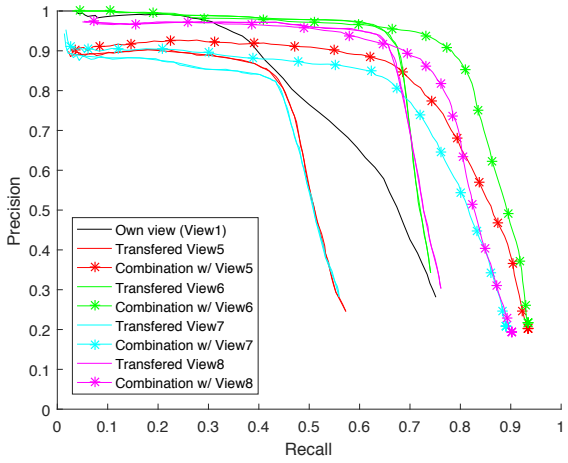

Fig. 5. AUC curves for the PETS2009 evaluation.

A detection is considered true if $rd \leq 0.5$ (corresponding to a deviation up to 25% of the true object size) and cover and overlap are both above 50%.

The integrated Average Precision is generally used to summarize the algorithm performance in a single value, represented geometrically as the area under the PR curve (AUC).

4. EXPERIMENTAL RESULTS

This section evaluates the proposed technique for information transference and combination between cameras. First, in order to verify that the proposed technique works correctly on detections with different appearance and aspect ratio, the evaluation on the dataset WUds is made. This is performed by detecting people with a standing person model and with a wheelchair user model, transferring the information to the other camera and evaluating, in both cameras, the own detections, the detection transferred from the other camera and the information combination from both cameras. Figure 4 shows the precision-recall curves and Table 2 (WUds column) shows the AUC value for each curve. The detections obtained by the evaluated cameras are better than the detections transferred from the other camera and the information combination improves the results of each camera separately. This scenario is relatively simple, since there are not too many people in the scene and therefore the results are relatively good in all the three cases, but it shows that the proposed technique works for detection of objects with different aspect ratio and an improvement is achieved by the proposed information combination.

The proposed technique is evaluated also on the PETS2009 sequences. As discussed in the dataset subsection (3.2.2), in this case the only available ground truth is from view 1, so all evaluations are performed with respect to this camera. Figure 5 shows the precision-recall curves and Table 2 (PETS2009 columns) shows the AUC value for each curve. In this case, the improvement obtained is much more remarkable than in the previous dataset. The camera facing the evaluated viewpoint is view 8: being closer to the monitored zone than view 1, it performs better than it, and the combination of both improves the final result score. The best result is obtained by combining view 6 with view 1: view 6 has a similar orientation to view 1 and, being closer to the scene, it gets the best individual score. Despite having the same orientation, the combination of these two cameras

		WUds	PETS2009			
Own view		0.733	0.619 (View1)			
Transferred view		0.715	View5 0.463	View6 0.706	View7 0.448	View8 0.701
Combination		**0.769**	0.775	**0.860**	0.720	0.795
Improvement ($\Delta\%$)	vs. own	4,7	20,1	28,0	14,0	22,1
	vs. transferred	7,0	40,3	17,9	37,8	11,8

Table 2. AUC Results for evaluation of PETS2009 and WUds.

obtains the best AUC of all curves as they record the same area from different distances. This is due to the small viewpoint change, which limits the possible errors when translating the detections. View 5 and view 7 are very similar to each other, as they correspond to the side views, one on each side with respect to view 1. Therefore, their results are very similar to each other, being slightly better the behavior of view 5 and its subsequent combination with view 1. It should be noted that in all camera combinations, the results obtained are better than those of any camera separately.

5. CONCLUSIONS

This paper presents a multi-camera system that adds contextual information of the scene to a people detector in order to properly combine information from the different cameras what results in an improvement of the performance of each of them independently. The results obtained after the evaluation of the system confirm the initial hypothesis, obtaining improvements in the detections performance by combining the cameras information in the two evaluated scenarios. The proposed method works for different people aspect ratio (standing, sitting, etc.) and for any orientation between the different cameras thanks to the proposed volumetric assumption.

As future work, it is proposed to use and combine other detectors (e.g the ones cited in subsection 3.1) to try to improve the system performance and to develop other elaborated information combination methods, for example by weighing the contribution of each camera to the combined detection as a function of distance between the camera and the object/person, since the closest detections to a camera are usually more accurate and have a greater confidence.

6. REFERENCES

[1] Á. García-Martín and J. M. Martínez, "People detection in surveillance: classification and evaluation," *IET Computer Vision*, vol. 9, no. 5, pp. 779–788, 2015.

[2] Xiaogang Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognition Letters*, vol. 34, no. 1, pp. 3–19, 2013.

[3] T. T. Santos and C. H. Morimoto, "People detection under occlusion in multiple camera views," *Proc. of Brazilian Symposium on Computer Graphics and Image Processing*, pp. 53–60, 2008.

[4] Kyungnam Kim and Larry S. Davis, *Multi-camera Tracking and Segmentation of Occluded People on Ground Plane Using Search-Guided Particle Filtering*, pp. 98–109, Springer Berlin Heidelberg, 2006.

[5] Thiago T. Santos and Carlos H. Morimoto, "Multiple camera people detection and tracking using support integration," *Pattern Recognition Letters*, vol. 32, no. 1, pp. 47–55, 2011.

[6] J. Black, T. Ellis, and P. Rosin, "Multi view image surveillance and tracking," *Proc. of Workshop on Motion and Video Computing*, pp. 169–174, 2002.

[7] P. Kilambi, E. Ribnick, A. J. Joshi, O. Masoud, and N. Papanikolopoulos, "Estimating pedestrian counts in groups," *Computer Vision and Image Understanding*, vol. 110, pp. 43–59, 2008.

[8] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," *Proc. of Computer Vision and Pattern Recognition*, pp. 878–885, 2005.

[9] 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*, vol. 32, no. 9, pp. 1627–1645, 2010.

[10] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 580–587, 2013.

[11] Ross B. Girshick, "Fast R-CNN," *Proc. of International Conference on Computer Vision*, pp. 1440–1448, 2015.

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 91–99, 2015.

[13] A. Milan, S. Roth, and K. Schindler, "Continuous energy minimization for multitarget tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 58–72, 2014.

[14] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *International Journal of Computer Vision*, vol. 77, pp. 259–289, 2008.