MULTI-VIEW OBJECT AND HUMAN BODY PART DETECTION UTILIZING 3D SCENE INFORMATION

Georgios Sfiris, Nikolaos Nikolaidis, Ioannis Pitas

Informatics and Telematics Institute, CERTH, Greece
Department of Informatics, Aristotle University of Thessaloniki, Greece

ABSTRACT

The aim of this paper is to present a new method for multi-view object or human body (or body part) detection. The basic idea consists of using a single view detector in every view of a scene captured by multiple cameras and then combining the results using the 3D information of the scene. The method can improve the results of the single view detector, while also localizing the object/human in the 3D space. This results in a robust way for rejecting the false detections, amending the missed detections and associating the results of the single view detector across views.

1. INTRODUCTION

The successful detection of an object's (or human's) location in images or videos has many applications that include tracking, object/human recognition, activity recognition, surveillance and robot vision. Numerous object or human body (or body part) detectors that operate on a single image or single view video have been developed in the past few years, having a fair success rate. However the issue of human or object detection in a convergent multiple camera environment has been scarcely touched upon, although such algorithms might find important applications in stereoscopic cinema or TV production and post production such as providing useful information for matting initialization or camera calibration (by excluding moving humans from the procedure). Obviously the existence of multi-view information is expected to lead to improved detection results, but most existing methods do not rely only to this information in order to achieve a 3D detection by matching detections across views. In [1] the authors make use of color histograms to associate data across different views. [2] represents human body parts such as upper and lower arm by line segments and matches only such structures across views. In [3] the 3D detection of human body is trivial since the scene captured is a football field providing for easy background extraction and a ground plane limiting the 3D search space. Calibration information is used in [4] to match the detections across views, but the utilized ray intersection technique requires a very accurate single view detector.

In this paper we propose a novel multi-view detection method that utilizes a single view detector to locate objects/humans in each one of the multiple views of the same scene obtained through calibrated cameras and uses this information along with the calibration information in order to locate the object in the 3D space and also improve the detection results in each view by eliminating false positive and false negative detections and associating detected objects, bodies or body parts (e.g. faces) across views.

The basic idea behind the proposed method is the following. Lets assume that we have an object in a scene and a number of images of this scene, some of which depict the object in question. We assume that the images have been obtained by a set of convergent calibrated synchronized cameras. Then, for every image there exists a non-linear transformation that relates the 3D coordinates of the object to the 2D coordinates of the object's projection on the image plane. These transformations provide a unique way to fuse the object location information in every image derived by the application of a certain detector, thus minimizing the effects of occlusion, providing 3D estimates of the object's location, and improving the accuracy and robustness of the 2D and 3D location estimation. At least two or more cameras, calibrated with respect to a common coordinate system should be available. An overview of the system is proposed in section 2. The details of the proposed method as well as the experimental results are provided in sections 3-6. Conclusions follow.

2. SYSTEM OVERVIEW

The procedure followed by the proposed multi-view object and human body part detector can be summarized in the following steps:

- 2D Detection. Having a set of images depicting a scene from different views, we use a single view detector to locate objects, human bodies or body parts, resulting in correct or false detections.

The research leading to these results has received funding from the European Communitys Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 211471 (i3DPost).

- Voting. By back projecting each detected instance of the objects, human bodies or body parts on the 3D space utilizing camera calibration information, we find the 3D regions where the created volumes intersect each other. Using a voting approach we find which of these regions correspond to the scene entities. Thus each selected 3D region corresponds to a single entity. A set of detected instances of this entity on each view is also associated with the 3D region. This step allows us to reject false detections in the various views.

- 3D Detection. In this step information from the voting step is used to associate detections across views so that all associated detections correspond to the same entity and also to detect the entity in views that the single view detector failed to do so.

3. 2D DETECTION

Without loss of generality, lets assume that there is one object in the scene and a set of cameras each capturing one image of the scene (Fig. 1(a)). In the 2D space of each image there exists an instance of the object, created by projecting every point of the object on the image. To locate the instance of the object in the 2D space of each image we use a single view detector (object, human body or body part detector). The detector may result in finding the instance of the object, missing it (false negatives), or finding false instances of the object (Fig. 2) i.e. parts of the image that are incorrectly detected as instances of the object (false positives).

The results of the detector are given in the form of bounding boxes (BBs) containing the object's instance. A BB returned by the detector is denoted by b, a set of BBs is denoted by B, b_{ij} is the j-th BB returned by the detector for the i-th camera and N_i is the total number of BBs returned for that camera. Our aim is a) to characterize each BB as being a valid or an invalid detection of the object b) generate BB in views that erroneously or due to occlusion lack them and c) associate BBs that correspond to the same object across views.

4. VOTING

We split the scene's volume thus creating a 3D grid of voxels (in line with the 2D grid of pixels in an image). Each voxel is denoted by v_{ijk}, a set of voxels is denoted by V and $\boldsymbol{M}(v_{ijk})$ is the center of voxel v_{ijk}. We define a discrete scalar field $F(v_{ijk})$ in the 3D space of the scene by projecting every voxel centroid on every camera and checking if the projection lies inside a detector-generated BB or not:

$$F(v_{ijk}) = \frac{\sum_{k=0}^{K-1} f_k(\boldsymbol{P}_k \boldsymbol{M}(v_{ijk}))}{K} \quad (1)$$

$$f_k(\boldsymbol{m}) = \begin{cases} 1, & \text{if } \boldsymbol{m} \in S_k, \\ 0, & \text{if } \boldsymbol{m} \notin S_k. \end{cases} \quad (2)$$

$$S_k = \bigcup_{0 \le j < N_k} b_{kj} \quad (3)$$

where $\boldsymbol{P}_k$ is the k-th camera's projection matrix and K is the number of cameras. In a sense, this field can be considered as the probability field of the object to exist in the point $\boldsymbol{M}$ of the 3D space (Fig. 1(b)). Then we reject every voxel with a low probability and create a set of "valid" voxels:

$$V' = \{v_{ijk} : F(v_{ijk}) > \alpha\} \quad (4)$$

where α is a properly selected threshold. By projecting every "valid" voxel $v \in V'$ on the i-th camera we create a region of high object existence probability denoted by H_i. Then, for every camera we reject BBs that have a small overlap with H and create a new set of BBs:

$$B'_k = \left\{ b_{kj} : \frac{A(b_{kj} \cap H_k)}{A(b_{kj})} > \beta \right\}, 0 \le k < K-1 \quad (5)$$

where $A(S)$ is the area of region S and β is a properly selected threshold. Having created a new set of BBs we create a new discrete probability field using (1) and repeatedly apply this method until no BB is rejected.

5. 3D DETECTION

5.1. Associating BBs across views

Once we have selected the "valid" BBs we must associate the BBs that correspond to the same object across different cameras. To achieve this we use the data stored in the "valid" voxels of the scene. Each "valid" voxel has been potentially voted by a specific BB from every camera. Thus this voxel associates a set of BBs from different views and contributes a vote for this set:

$$Set = \{J_0, J_1, \ldots, J_{K-1}\} \quad (6)$$

$$J_i = \begin{cases} j, & \text{if } b_{ij} \text{ is selected for the } i\text{-th camera} \\ -1, & \text{if no BB is selected for the } i\text{-th camera} \end{cases} \quad (7)$$

By counting these votes we create a list containing these sets in an ascending order according to the number of votes (Fig. 3):

$$List = \begin{cases} Set_0 & Votes_0 \\ Set_1 & Votes_1 \\ \vdots & \\ Set_{L-1} & Votes_{L-1} \end{cases} \quad (8)$$

$$Votes_0 \ge Votes_1 \ge \cdots \ge Votes_L \quad (9)$$

where L is the number of sets. Note that two or more sets may contain the same BB from the same camera (Fig. 3(b)).

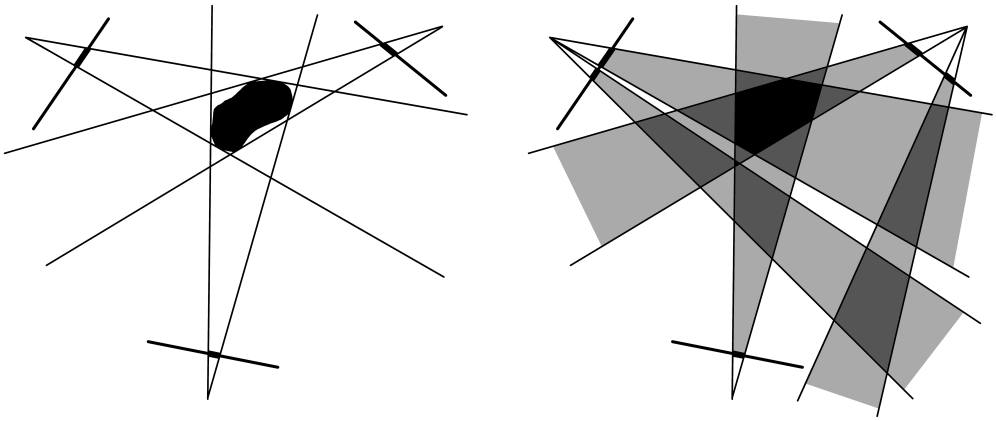

Fig. 1. (a) Three cameras capturing a scene and depicting one instance of the object in their corresponding images. (b) The probability field created by back-projecting every bounding box returned by the object detector. Darker areas denote higher probability of the object to exist in that part of the 3D space.

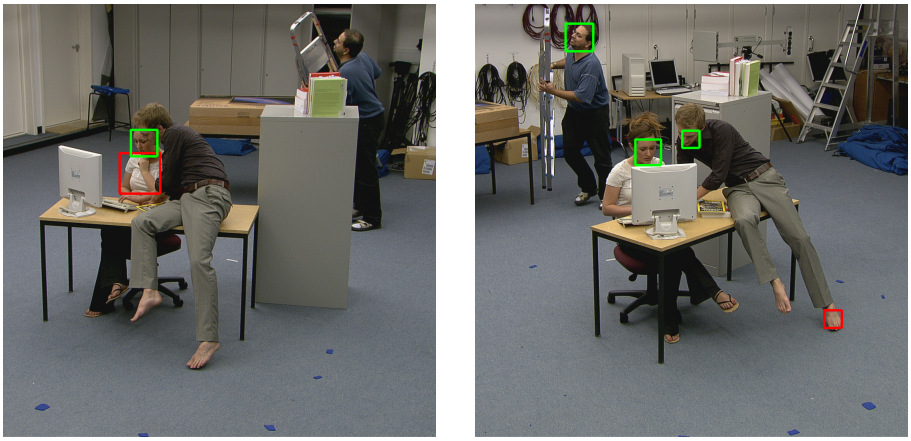

Fig. 2. The results of a face detector on two views of the same scene. Correct detections are marked with green color while false detections are marked with red color.

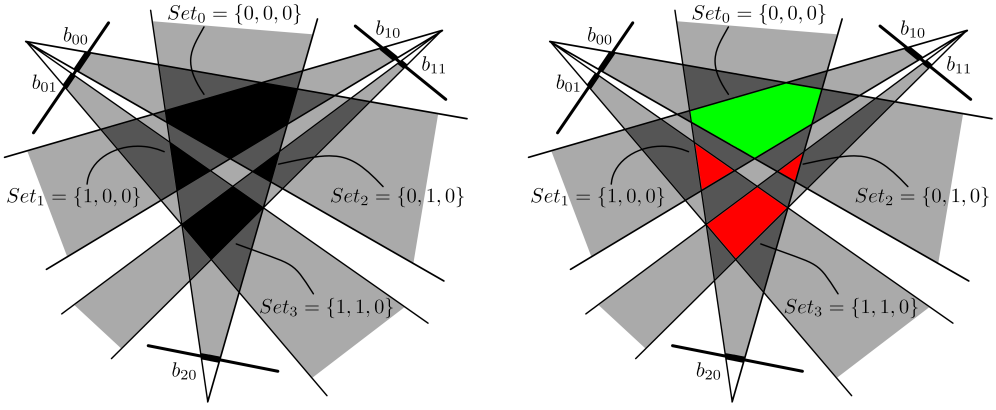

Fig. 3. (a) Set of bounding boxes voted by the corresponding retained areas (black color). The number of votes is proportional to the size of the area. (b) The valid set of bounding boxes (voted by voxels in the green color area) and the rejected sets (red color).

We consider that Set_0 is a valid set of BBs that point to the same (existing) object and add this set to a new list, $List'$, that contains only the valid sets. Then for all remaining sets Set_i, $0 < i \leq L-1$, in $List$ we check if any BB of Set_i is used already in $List'$. If not, we add Set_i in $List'$, else we remove the common BBs from Set_i and, provided that the remaining BBs in Set_i are enough in number, we add Set_i in $List'$. At the end, $List'$ will contain all valid sets of (now associated) BBs, each set corresponding to a specific object of the scene.

5.2. Rectification

Once we have found the sets of BBs that correspond to a specific object we can create a BB for every view that has no associated BB for this object. By back projecting a set of BBs we create a volume that corresponds to a specific object or human. By projecting this volume on views that have no BB we create a surface that potentially contains the object projection. We can create a BB for that view, using the volume's or surface's centroid or the area of that surface.

6. RESULTS

To test this method a variation of the face detector proposed in [5], trained to detect both frontal and profile faces, was used. In the utilized variation a skin color detector was added in order to improve the results. In more detail, the pixels' color in the facial BBs returned by the frontal/profile detector in [5] were checked against a range of skin-like colors. BBs that contained a small percentage of skin-like colored pixels were rejected. The method was tested on four image sequences depicting people standing, sitting, walking, running, lying, carrying thing, undressing, being occluded etc. The scenes were captured using eight convergent calibrated HD cameras having a resolution of 1920×1080 pixels and a capturing rate of 25 frames/second. The method was applied to every fifth frame of each multi-view sequence and was evaluated against manually constructed ground truth (head BBs). Table 1 represents the improvement brought to the 2D detection by using the proposed approach. It is obvious that the method significantly increases correct detections and reduces false positives and negatives. A correct detection was declared when the ground truth and algorithm generated BBs had an overlap in excess of 70%.

It is worth noting that having the 3D location of the detected head (which is a byproduct of the method) is a powerful information that can be used to improve the results of the detection. For example if we have information about the 3D location of the floor or the ceiling planes, we can reject any head detection close to these planes. In addition we can make use of a 3D tracker to improve the results using temporal information.

	Correct	False Negatives	False Positives
Before	$957(40\%)$	$1414(59\%)$	661
After	$1625(68\%)$	$746(31\%)$	373
Difference	$+668(28\%)$	$-668(28\%)$	-288

Table 1. Detection results on all views before (i.e. by using the face detector in each view separately) and after the application of the proposed method for four image sequences of a man standing, undressing and lying (Fig. 2), a man and a woman fighting with pillows, a man and a woman sitting and a man walking and carrying a ladder, a man standing in a counterfeit prison.

7. CONCLUSION

We have proposed a novel simple to use multi-view detection method. This method utilizes a single view detector to locate objects in each one of the multiple views of the same scene obtained through calibrated cameras and uses the calibration information to match the detections across views. Any object, body or body part single view detector can be used to feed the multi-view detector, resulting in an object, body or body part multi-view detector respectively. The method improves the detection results in each view by eliminating false positive and false negative detections and associating detected objects, bodies or body parts (e.g. faces) across views.

8. REFERENCES

[1] Luca Marchesotti, Gianni Vernazza, and Carlo S. Regazzoni, "A multicamera fusion framework for multiple occluding objects tracking in intelligent monitoring and sport viewing applications," in *IEEE ICIP*, 2004, pp. 1033–1036.

[2] Abhinav Gupta, Anurag Mittal, and Larry S. Davis, "Constraint integration for multiview pose estimation of humans with self-occlusions," in *Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06)*, Washington, DC, USA, 2006, pp. 900–907.

[3] Ming Xu, James Orwell, and Graeme Jones, "Tracking football players with multiple cameras," in *IEEE ICIP*, 2004, pp. 2909–2912.

[4] James Black, Tim Ellis, and Paul Rosin, "Multi view image surveillance and tracking," in *MOTION '02: Proceedings of the Workshop on Motion and Video Computing*, Washington, DC, USA, 2002, pp. 169–174.

[5] Paul Viola and Michael J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.