A SPARSITY CONSTRAINED INVERSE PROBLEM TO LOCATE PEOPLE IN A NETWORK OF CAMERAS

Alexandre Alahi[1], *Yannick Boursier*[1], *Laurent Jacques*[1,2] , *Pierre Vandergheynst*[1]

[1]Institute of Electrical Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland

[2]Communications and Remote Sensing Laboratory, Université catholique de Louvain (UCL), B-1348 Louvain-la-Neuve, Belgium

ABSTRACT

A novel approach is presented to locate dense crowd of people in a network of fixed cameras given the severely degraded background subtracted silhouettes. The problem is formulated as a sparsity constrained inverse problem using an adaptive dictionary constructed on-line. The framework has no constraint on the number of cameras neither on the surface to be monitored. Even with a single camera, partially occluded and grouped people are correctly detected and segmented. Qualitative results are presented in indoor and outdoor scenes.

Index Terms— Inverse problem, Sparsity, People detection

1. INTRODUCTION

Vision-based human detection has been of interest to the research community for the past decades. Many applications such as surveillance, security, medical, and safety are interested in detecting and tracking people in crowded scene. Counting the number of pedestrians, detecting abnormal behaviors, analyzing behavior, localizing people to avoid dangerous traffic situations rely on an accurate detection algorithm. The performance of the algorithms depends on the sensing modality. In this work, fixed cameras are of interest. Scenes requirement for monitoring are typically larger than the field of view of a single camera. The presence of multiple cameras is necessary. It can solve problems arising with monocular vision such as occlusions.

Porikli in [1] presents a survey on object detection and tracking methods with a fixed camera. Given a fixed camera, objects can be detected by modeling the background and tracking become simply an object correspondence across frames. Typically, the work of Stauffer and Grimson [2] can be used to extract the foreground pixels. Each pixel is modeled as a mixture of Gaussians with an on-line approximation for the update. Then, inter-frame tracking algorithms such as the one presented by Avidan in [3] can be used to track people across frames. He considers the tracking as a binary classification problem. However, those algorithms are less efficient to detect people in a dense environment. A group of people is not correctly segmented due to their mutual occlusions.

In order to deal with occlusions, the output of several cameras are fused to detect the objects of interest. To be robust to variability in appearance between the views, coordinates of the detected objects in a common reference (*e.g.* ground plane) is estimated. The unique 'world' coordinates, *i.e.* the coordinate of the object on the ground plane, is given by a planar homography. The planar homography is a 3×3 matrix transformation obtained by matching at least four points from two different coordinates. Most of the systems compute the homographies at initial calibration step [4]. After projecting all detected objects into a common reference, Mueller *et al.* in [4] mark with same label the nearest object with the same size and center of gravity. Orwell *et al.* in [5] and Caspi *et al.* in [6] match objects by fusing the estimated trajectories obtained by each camera. However such approaches are not fully taking advantage of the multi-view infrastructure.

Khan and Shah in [7] present an approach to track people in crowded scene with multiple cameras located at a distance close to head level. They pay attention to extract the feet region of the foreground people. Each point of the foreground likelihood (background subtracted silhouettes) from all views is mapped to the ground plane given a planar homography. Multiplying the mapped points segment the pixel pertaining to the feet of the people. Their approach can not be applied to an object viewed by one camera. In addition, a poor foreground segmentation - people detected with their shadow - affects the performance of their system.

Fleuret *et al.* in [8] show that they can combine a generative model with dynamic programming to accurately track people across multiple cameras. The cameras are also setup at head level. They develop a mathematical framework to estimate the probabilities of occupancy of the ground plane at each time frame with a dynamic programming to track people over time. They approximate the occupancy probabilities as the marginals of a product law minimizing the Kullback-Leibler divergence from the true conditional posterior distribution (referred to as Fixed Point Probability Field algorithm). These probabilities are combined with a basic color

Y. B. is a Postdoctoral Researcher funded by the APIDIS European Project. L. J. is a Postdoctoral Researcher of the Belgian National Science Foundation (F.R.S.-FNRS).

and motion model. As in [7], their approach can not be applied to an object viewed by one camera. In addition, the complexity cost of their algorithm depends on the number of ground plane points to be evaluated leading to a limited area to be monitored.

More recently, Reddy *et al.* in [9] use compress sensing theory to track people in a multi-view setup. They use the sparsity present in the background subtracted silhouettes extracted from the cameras. Their sparsity constraint depends on the distance of the objects to the cameras. Objects close to the cameras will generate large background subtracted silhouettes with poor sparsity. To accurately estimate the position of the objects on the ground plane multiple cameras are needed. Also, the complexity cost of their algorithm depends on the number of ground plane points, the grid size, to be evaluated.

In this work, a novel approach is presented to locate dense and occluded people with several fixed cameras given the severely degraded background subtracted silhouettes. The framework scales to any number of cameras. A single camera can also be used. The proposed algorithm is based on a sparse approximation of the location points given an adaptive dictionary constructed on-line. The sparsity constraint in this work is not on the observations but on the desired solution. Even when a dense crowd of people is present in a scene, they occupy sparse locations on the ground plane. Whereas previous works use fixed ground plane grid points, in this work a sparse grid is proposed reducing consequently the complexity cost. Each atom of the dictionary is generated given the sparse grid points.

The rest of the paper is structured as follows: after formulating the problem, Section 3 describes the dictionary and its atoms. Then, the proposed on-line adaptive dictionary construction is presented in Section 4. Finally, the paper ends with qualitative results.

2. PROBLEM FORMULATION

The objective of this paper is to deduce the ground plane points (or *grid of occupancy*) occupied by the people present in the scene given the background subtracted silhouettes provided by a set of C calibrated cameras. Mathematically, at a given time, each camera is the source of a binary silhouette image $y_i \in \{0,1\}^{M_i}$, where $M_i \in \mathbb{N}$ is the number of pixels of each camera indexed by $1 \leq c \leq C$. Stacking all these vectors, we know of the observed scene the Multi-Silhouette Vector (MSV) $y = (y_1^T, \cdots, y_N^T)^T \in \{0,1\}^M$, with $M = \sum_i M_i$.

Let us discretize the observed ground in N subareas, so that the grid of occupancy of people on the ground is represented by the binary vector[1] $x \in \{0,1\}^N$. For simplicity, we assume that one observed person is exactly supported by one subarea of this grid.

[1]The grid is of course bidimensional but we represent it as a vector to simplify the notations.

It is clear that any configuration of x will correspond to a particular configuration of silhouettes in y. For instance, if x contains only one non-zero component, all y_i will contain only one silhouette (i.e. a connex area of non-zero pixels) with size and location related to the particular projective geometry combining the scene and the cameras.

Our inverse problem is thus to find x from y. However, the number of observations M may be much smaller than the grid size N. In addition, the vector y is binary. It does not contain any information about possible occlusion between persons, and the background subtracting methods leading to the silhouette definition are severely degraded.

We decide therefore to assume the grid x sparse, i.e. it has only few non-zero coefficients.

The generative (forward) model that associates to x a certain configuration of silhouette in y is assumed linear (at first order). In other words, we obtain it from a *dictionary* $D \in \mathbb{R}^{M \times N}$. As explained in Section 3, each column (or atom) of D transforms a non-zero component of x to the corresponding silhouette generated by a person in the image plane of the cameras.

The recovery of the grid of occupancy from y is thus solved by the following program:

$$\arg\min_x \|x\|_0 \text{ s.t. } \|y - Dx\|_2 < \varepsilon \tag{1}$$

with ε is the desired residual error

The ℓ_0 norm is the appropriate measure of the sparsity since it counts the number of non-zero elements. However, minimizing ℓ_0 norm is NP-hard and is prohibitive in high dimension. Hence, we approximate the ℓ_0 norm with a reweighted ℓ_1 norm where the weights used for the next iteration are computed from the value of the current solution as introduced by Candes *et al.* in [10]:

$$x^{(l+1)} = \arg\min_{u \in R^N} \|W^{(l)} u\|_1 \text{ s.t. } \|y - Du\|_2 < \varepsilon, \tag{2}$$

$$W^{(l+1)} = \text{diag}\left(\frac{1}{|x_1^{(l)}|+\eta}, \cdots, \frac{1}{|x_N^{(l)}|+\eta}\right), \tag{3}$$

with $W^0 = \text{Id}$.

The parameter η is added to assure stability and guarantees that a zero-valued component in x does not strictly prohibit a nonzero estimate at the next iteration.

In this paper, we solve each weighted iteration of the reweighted process by the method of operator splitting and proximal methods [11, 12].

3. DICTIONARY CONSTRUCTION

The dictionary D is made of atoms modeling the presence of a single person at a given location. Each atom of the dictionary represents the ideal silhouette generated by a person in the image plane of each camera, i.e the MSV. Planar homographies computed at calibration step are used to map points

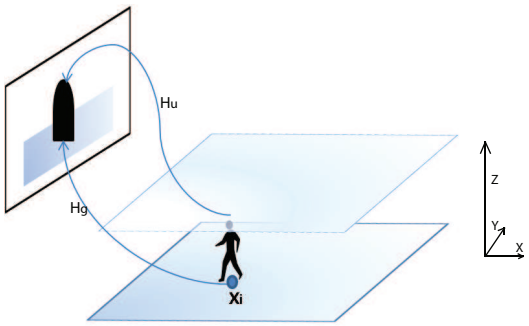

Fig. 1. To each point x_i corresponds a silhouette modeling the presence of a person in a camera view

on the ground plane to the image plane of each camera:

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} H_{11} & H_{12} & H_{13} \\ H_{21} & H_{22} & H_{23} \\ H_{31} & H_{32} & H_{33} \end{pmatrix} \begin{pmatrix} X \\ Y \\ 1 \end{pmatrix}$$

where H_{ij} is an element of the homography matrix, and (x, y) are the pixel location in the image plane of a camera corresponding to the ground plane point (X, Y).

$$x = \frac{XH_{11} + YH_{12} + H_{13}}{XH_{31} + YH_{32} + H_{33}}$$

and

$$y = \frac{XH_{21} + YH_{22} + H_{23}}{XH_{31} + YH_{32} + H_{33}}$$

The homography matrix of the plane parallel to the ground plane at a height similar to the average pedestrian (1m70) is also computed to generate the silhouettes observed in the image plane of the cameras (see figure 1). Vanishing line and points are used to compute such homography [13]:

Given the ground plane homography H_g,

$$H_g = \begin{pmatrix} H_1 & H_2 & H_3 \\ | & | & | \end{pmatrix}$$

The upper plane translated by z units is simply:

$$H_u = \begin{pmatrix} H_1 & H_2 & \alpha.z.v_z + H_3 \\ | & | & | \end{pmatrix}$$

where α is a scalar factor, v_z is the vanishing point for the Z direction (vertical lines).

To cope with the various poses and shapes a person can generate in a camera view, a half elliptical shape is used to approximate the ideal silhouette of a person. Figure 2 illustrates an example of severely degraded background subtracted silhouettes and the silhouettes used to model their presence at the same locations.

The silhouettes generated by a camera c lead to the dictionary D_c. The product $D_c x$ corresponds to the synthetic image

(a) Camera View c

(b) Degraded background subtracted silhouettes y_c

(c) $\hat{y}_c = D_c x$

Fig. 2. Illustration of the atoms modeling the given background subtracted silhouettes.

$\hat{y}_c$ illustrated in figure 2. Each column of the dictionary D_c is the image of a single silhouette reshaped as a vector. There are as many columns as ground plane points. When several cameras are observing a scene, D is simply the stack of the D_c:

$$D = \begin{pmatrix} D_1 \\ D_2 \\ \vdots \\ D_n \end{pmatrix}$$

where n is the number of cameras. Also, as presented in the previous section, the Multi-Silhouette Vector (MSV) is stacked as follows:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}$$

With such formalization, there is no constraint on the number of cameras to use.

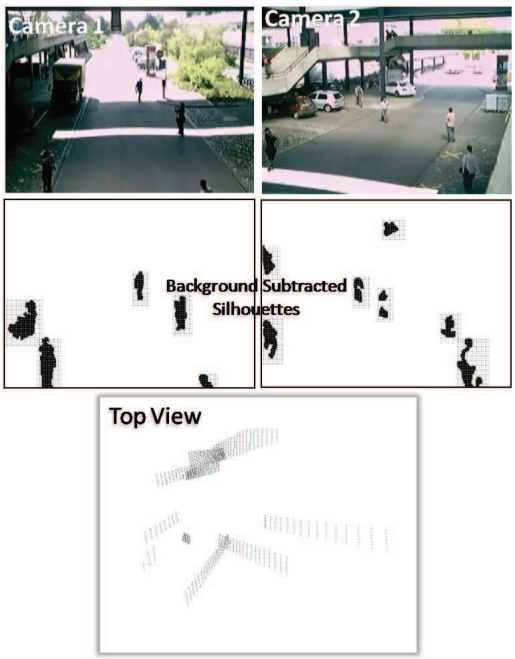

Fig. 3. Only gray points in the top view are forming x. Those points are the potential locations of the people.

4. ON-LINE ADAPTIVE DICTIONARY CONSTRUCTION

The complexity cost depends on the number N of ground plane points to locate as occupied or not. Previous works considered a fixed number of points regardless the geometry of the scene and the sparsity of the people present in the scene. In addition, the ground plane points were sampled uniformly without considering the resolution of the observations. In this work, the ground plane points are sampled based on the resolution of the observations.

Two different ground plane points can correspond to the same pixel in the image plane of a camera. A translation of one pixel in the image plane can be equivalent to a translation of a few meters on the ground plane for far away regions, just as a translation of a few centimeters on the ground plane can correspond to a shift of several pixels for closer regions. It depends on the resolution of the camera and the distance of the objects to the camera. Therefore, localization should be on a restricted number of ground plane points, called sample points.

The bounding box of the background subtracted silhouettes determines the sampling points to be used (see Figure 3). Each point is spaced by the "just noticeable pixel step". It represents the pixel accuracy to detect an object in the image plane. Given the planar homography H_g, each point of each camera is mapped to a ground plane point forming x.

5. EXPERIMENTS

Indoor and outdoor data sets have been used. Cameras are located at a height equivalent to the first and second floor of a building. The images are recorded at 25 fps with a resolution of 320×240. The APIDIS dataset[2] is also used to locate basket ball players in challenging interactions.

The background subtracted silhouettes extracted from those sequences are severely degraded. Only part of the people are extracted, their shadow is considered, and random false positives are generated. No ground truth is available for these data sets, thus no quantitative criteria can be computed. However, the visual quality of results, with a single camera (displayed in Figure 4) or two cameras (displayed in Figure 5) in the dataset, highlights the robustness of our algorithm to such noisy background subtracted silhouettes.

People are correctly located and the algorithm efficiently prevents from false detection and multiple detection for a same person, even for people that are partially occluded or that are grouped. The use of two cameras improves the accuracy of results, particularly in crowded scene with occlusions as the APIDIS dataset.

Figure 6 illustrates the increasing of performance obtained with the reweighted approach compared here with the basic ℓ_1 minimization that is deduced by replacing the ℓ_0 norm by the ℓ_1 norm in equation 1. The sparsity of the solution is clearly improved when solving the reweighted problem. The solution of the ℓ_1 problem suffers from multiple detection for the same basket ball players, whereas the reweighted approach removes these ambiguities.

Note that if the reweighted ℓ_1 minimization is not used, the detected locations are not sparse enough. Several locations are detected for a single pedestrian (see Figure 6).

Fleuret *et al.* in [8] use a rectangular silhouette to model people. Figure 7 illustrates the performance when a rectangular shape is used instead of a half elliptical. A full elliptical silhouette is also presented to show relevance of our half elliptical shape. People are better detected with the proposed silhouette.

Finally, the adaptive dictionary construction speeds-up the processing time radically and locate people more precisely since points do not need to fall in an uniform sampling grid. A few hundred of points are enough to detect a dense crowd of 12 people whereas previous works needed thousands of points in a wide scene.

6. CONCLUSIONS

Formulating the localization of people as a sparse constrained inverse problem empower challenging localizations. Grouped and occluded people can be detected given very noisy observations. The problem is efficiently, if only approximately,

[2]The dataset is publicly available at `http://www.apidis.org/Dataset/`

Fig. 5. Four examples where people are segmented and detected with two cameras.

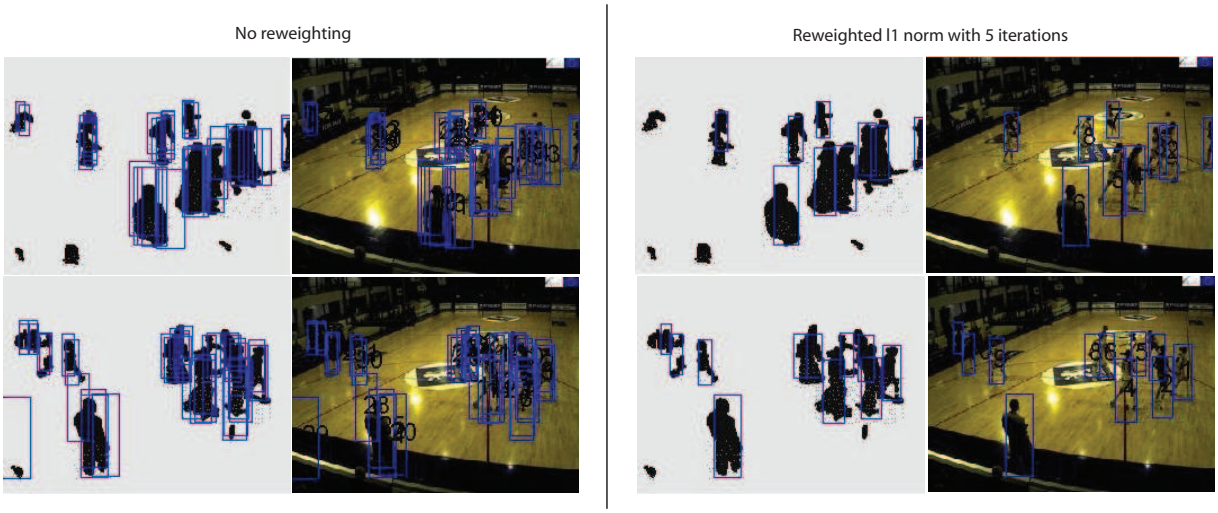

Fig. 6. Left hand-side presents the detected regions by a single camera without the reweighted ℓ_1 minimization. Right hand-side illustrates the detected regions with the reweighted minimization with five iterations only.

Fig. 4. Given a single camera and degraded background subtracted silhouettes, people are correctly segmented and detected

Fig. 7. Impact of various silhouettes to model people.

solved with an efficient re-weighted ℓ_1 strategy. Whereas previous works supposed several cameras, our approach works with a single camera and scales easily to any number of cameras. Finally, there is no constraint on the surface to be monitored. The proposed adaptive dictionary handles the sparsity of the problem also at the construction stage. Qualitative results have illustrated the strength of our approach on challenging scenarios. Further work will compare quantitatively the proposed approach with state-of-the-art multi-view systems. Mobile cameras can also be integrated to the system solving the optimization problem in a distributed fashion. Finally, the generative model that associates silhouettes to a ground plane point will be considered as non-linear to better deal with occlusions.

7. REFERENCES

[1] F. Porikli, "Achieving real-time object detection and tracking under extreme conditions," *Journal of Real-Time Image Processing*, vol. 1, no. 1, pp. 33–40, 2006.

[2] C. Stauffer and W.E.L. Grimson, "Adaptive background mixture models for real-time tracking," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 246–252, 1999.

[3] S. Avidan, "Ensemble tracking," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, pp. 261–271, 2007.

[4] K. Mueller, A. Smolic, M. Droese, P. Voigt, and T. Wienand, "Multi-texture modeling of 3d traffic scenes," *icme*, vol. 2, pp. 657–660, 2003.

[5] J. Orwell, S. Massey, P. Remagnino, D. Greenhill, and G. A. Jones, "A multi-agent framework for visual surveillance," in *ICIAP '99: Proceedings of the 10th International Conference on Image Analysis and Processing*, Washington, DC, USA, 1999, p. 1104, IEEE Computer Society.

[6] Y. Caspi, D. Simakov, and M. Irani, "Feature-based sequence-to-sequence matching," vol. 68, no. 1, pp. 53–64, June 2006.

[7] S.M. Khan and M. Shah, "A multiview approach to tracking people in crowded scenes using a planar homography constraint," 2006, pp. IV: 133–146.

[8] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multi-camera people tracking with a probabilistic occupancy map," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007.

[9] D. Reddy, A.C. Sankaranarayanan, V. Cevher, and R. Chellappa, "COMPRESSED SENSING FOR MULTI-VIEW TRACKING AND 3-D VOXEL RECONSTRUCTION," in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, 2008, pp. 221–224.

[10] EJ Candès, MB Wakin, and SP Boyd, "Enhancing sparsity by reweighting l 1," Tech. Rep., Tech. Rep., California Institute of Technol., 2007 [Online]. Available: http://www. acm. caltech. edu/ emmanuel/publications. html.

[11] P.L. Combettes, "Solving monotone inclusions via compositions of nonexpansive averaged operators," *Optimization*, vol. 53, no. 5, pp. 475–504, 2004.

[12] M.J. Fadili and J.-L. Starck, "Monotone operator splitting for fast sparse solutions of inverse problems," *SIAM Journal on Imaging Sciences*, 2009, submitted.

[13] S.M. Khan, P. Yan, and M. Shah, "A Homographic Framework for the Fusion of Multi-view Silhouettes," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 2007, pp. 1–8.