Hybrid Focal Stereo Networks for Pattern Analysis in Homogeneous Scenes

Emanuel Aldea[1,3] and Khurom H. Kiyani[2,3]

[1]Autonomous Systems Group, Université Paris Sud, France
[2]Communications and Signal Processing Group, Dept. of Electrical and Electronic Engineering, Imperial College London, UK
[3]AquaMed Research and Education, Doha, Qatar

Outline

1. Context
2. The proposed algorithm
3. Experimental results
4. Conclusions and future work

Outline

1. Context

2. The proposed algorithm

3. Experimental results

4. Conclusions and future work

The context of this work

Modelling high-density crowded scenes

- Understanding pedestrian dynamics at high densities
- Understanding how instabilities may build up
- Micro-analysis: in order to model the system, the particles (pedestrians) must be tracked individually

The context of this work

Modelling high-density crowded scenes

- Understanding pedestrian dynamics at high densities
- Understanding how instabilities may build up
- Micro-analysis: in order to model the system, the particles (pedestrians) must be tracked individually

The context of this work

Modelling high-density crowded scenes

- Understanding pedestrian dynamics at high densities
- Understanding how instabilities may build up
- Micro-analysis: in order to model the system, the particles (pedestrians) must be tracked individually

The context of this work

Modelling high-density crowded scenes

- Understanding pedestrian dynamics at high densities
- Understanding how instabilities may build up
- Micro-analysis: in order to model the system, the particles (pedestrians) must be tracked individually

The proposed strategy

- A common FOV multiple camera network, in order to cope with occlusion and clutter
- Redundancy also useful for filtering out spurious information (false detections, wrong associations)
- A compromise between FOV and resolution per pixel

The context of this work

Modelling high-density crowded scenes

- Understanding pedestrian dynamics at high densities
- Understanding how instabilities may build up
- Micro-analysis: in order to model the system, the particles (pedestrians) must be tracked individually

The proposed strategy

- A common FOV multiple camera network, in order to cope with occlusion and clutter
- Redundancy also useful for filtering out spurious information (false detections, wrong associations)
- A compromise between FOV and resolution per pixel

The context of this work

Modelling high-density crowded scenes

- Understanding pedestrian dynamics at high densities
- Understanding how instabilities may build up
- Micro-analysis: in order to model the system, the particles (pedestrians) must be tracked individually

The proposed strategy

- A common FOV multiple camera network, in order to cope with occlusion and clutter
- Redundancy also useful for filtering out spurious information (false detections, wrong associations)
- A compromise between FOV and resolution per pixel

The context of this work

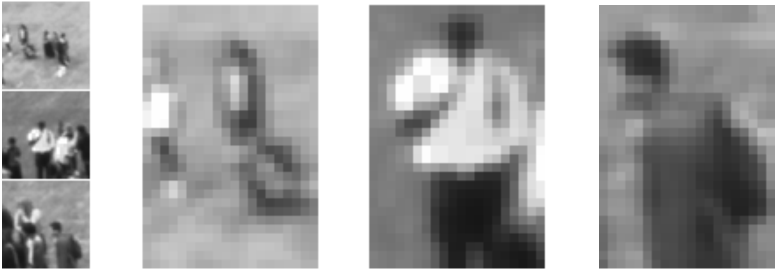
Modelling high-density crowded scenes

- Understanding pedestrian dynamics at high densities
- Understanding how instabilities may build up
- Micro-analysis: in order to model the system, the particles (pedestrians) must be tracked individually

The proposed strategy

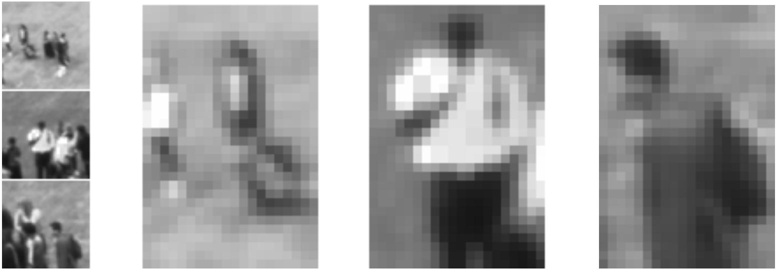
- A common FOV multiple camera network, in order to cope with occlusion and clutter
- Redundancy also useful for filtering out spurious information (false detections, wrong associations)
- A compromise between FOV and resolution per pixel

Wide FOV vs. narrow FOV



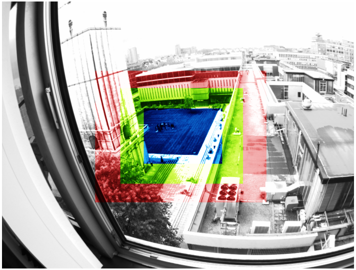
- Same sensor and position, three different focal lengths
- Long focal: required for retrieving details (head, body parts, bags etc.)
- Short focal: beneficial for accurate registration in a camera network
- By lacking salient features the narrow FOV is not able to estimate robustly or at all the relative pose between multiple cameras

Wide FOV vs. narrow FOV



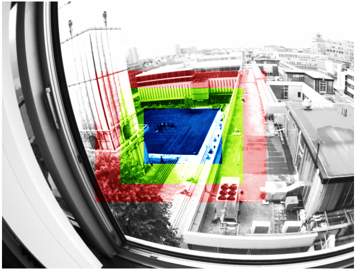
- Same sensor and position, three different focal lengths
- Long focal: required for retrieving details (head, body parts, bags etc.)
- Short focal: beneficial for accurate registration in a camera network
- By lacking salient features the narrow FOV is not able to estimate robustly or at all the relative pose between multiple cameras

Wide FOV vs. narrow FOV



- Same sensor and position, three different focal lengths
- Long focal: required for retrieving details (head, body parts, bags etc.)
- Short focal: beneficial for accurate registration in a camera network
- By lacking salient features the narrow FOV is not able to estimate robustly or at all the relative pose between multiple cameras

Wide FOV vs. narrow FOV



- Same sensor and position, three different focal lengths
- Long focal: required for retrieving details (head, body parts, bags etc.)
- Short focal: beneficial for accurate registration in a camera network
- By lacking salient features the narrow FOV is not able to estimate robustly or at all the relative pose between multiple cameras

Solutions for analysis and registration

Available solutions

- Motorized zoom lenses and PTZ cameras:
  - Re-estimation of intrinsic and distortion parameters
  - Simplifying assumptions about a subset of the varying parameters
  - Do not require precise reprojections
  - Still require constantly the presence of salient features
- Study of high-density crowds
- Study of different types of flows encountered in natural phenomena

Solutions for analysis and registration

Available solutions

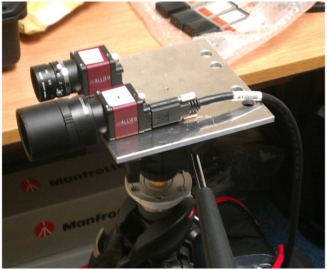
- Motorized zoom lenses and PTZ cameras:
  - Re-estimation of intrinsic and distortion parameters
  - Simplifying assumptions about a subset of the varying parameters
  - Do not require precise reprojections
  - Still require constantly the presence of salient features
- Study of high-density crowds
- Study of different types of flows encountered in natural phenomena

Solutions for analysis and registration

Available solutions

- Motorized zoom lenses and PTZ cameras:
  - Re-estimation of intrinsic and distortion parameters
  - Simplifying assumptions about a subset of the varying parameters
  - Do not require precise reprojections
  - Still require constantly the presence of salient features
- Study of high-density crowds
- Study of different types of flows encountered in natural phenomena

The proposed strategy



A hybrid stereo system

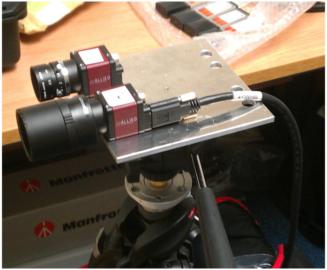
- Idea: use simultaneosly one camera for analysis and the other one for registration

✓ Accurate offline intrinsic calibration

✓ No salient features required in the FOV of the long focal camera

Related work

The proposed strategy



A hybrid stereo system

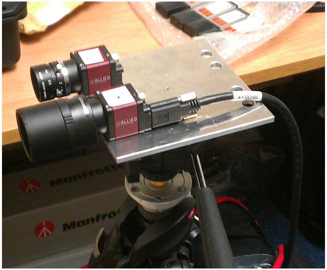
- Idea: use simultaneosly one camera for analysis and the other one for registration

✓ Accurate offline intrinsic calibration

✓ No salient features required in the FOV of the long focal camera

Related work

The proposed strategy



A hybrid stereo system

- Idea: use simultaneosly one camera for analysis and the other one for registration

✓ Accurate offline intrinsic calibration

✓ No salient features required in the FOV of the long focal camera

Related work

The proposed strategy

A hybrid stereo system

- Idea: use simultaneosly one camera for analysis and the other one for registration

✓ Accurate offline intrinsic calibration

✓ No salient features required in the FOV of the long focal camera

Related work

- the setup deployed by the STEREO solar observation mission
- more recently in robotics, using fisheye and perspective cameras
- a very recent paper[a] which proposed an AR binocular system

[a] Oskiper et al.: Augmented reality binoculars. ISMAR 2013

The proposed strategy

A hybrid stereo system

- Idea: use simultaneosly one camera for analysis and the other one for registration

✓ Accurate offline intrinsic calibration

✓ No salient features required in the FOV of the long focal camera

Related work

- the setup deployed by the STEREO solar observation mission
- more recently in robotics, using fisheye and perspective cameras
- a very recent paper[a] which proposed an AR binocular system

[a] Oskiper et al.: Augmented reality binoculars. ISMAR 2013

The proposed strategy

A hybrid stereo system

- Idea: use simultaneosly one camera for analysis and the other one for registration

✓ Accurate offline intrinsic calibration

✓ No salient features required in the FOV of the long focal camera

Related work

- the setup deployed by the STEREO solar observation mission
- more recently in robotics, using fisheye and perspective cameras
- a very recent paper[a] which proposed an AR binocular system

[a] Oskiper et al.: Augmented reality binoculars. ISMAR 2013

Outline

1. Context
2. The proposed algorithm
3. Experimental results
4. Conclusions and future work

The formulation of the problem

The variables involved

- N hybrid stereo rigs, i^{th} rig $= \{C_i^s, C_i^l\}$
- Objective: align accurately the cameras $\{C_1^s, C_2^s, \ldots, C_N^s\}$
- $\mathbf{E}_i^{sl}, \mathbf{E}_{ji}^l, \mathbf{E}_{ji}^s$: transforms within the same rig, and between rigs respectively
- $\mathbf{K}_i^s$, $\mathbf{K}_i^l$: intrinsic parameters, considered known
- $\mathbf{E}_i^{sl}$: related to rig extrinsic parameters, considered known

Two main steps

1. estimating $\mathbf{E}_{ji}^l$ using scene structure
2. transfering $\mathbf{E}_{ji}^l$ information to $\mathbf{E}_{ji}^s$

The main steps

Estimation of $\mathbf{E}^l_{ji}$

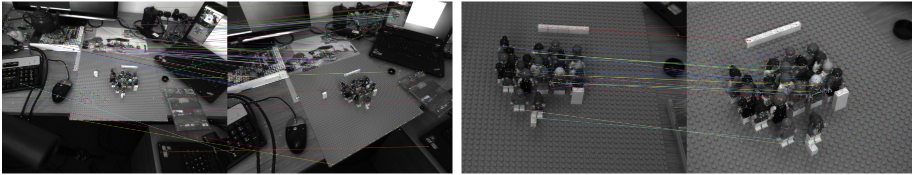
- $\mathbf{F}^l_{ji}$: SIFT detection and matching, normalized 8-point algorithm and RANSAC
- Matching step: we filter based on the uniqueness assumption and on married matching
- Decomposition of $\mathbf{F}^l_{ji}$ to get an estimation $\tilde{\mathbf{E}}^l_{ji}$
- Based on the inliers, we triangulate a set of 3D points $\tilde{\mathbf{X}}_{ji}$
- We use $\tilde{\mathbf{X}}_{ji}$ and $\tilde{\mathbf{E}}^l_{ji}$ in a BA procedure and we get $\hat{\mathbf{E}}^l_{ji}$

Global alignment of analysis cameras

$$\mathbf{E}^s_{ji} = \mathbf{E}^{sl}_j \hat{\mathbf{E}}^l_j \left(\mathbf{E}^{sl}_i \hat{\mathbf{E}}^l_i \right)^{-1}$$

where $\hat{\mathbf{E}}^l_i, \hat{\mathbf{E}}^l_j$ are expressed in a fixed reference system.

A small scale scenario



a) b)

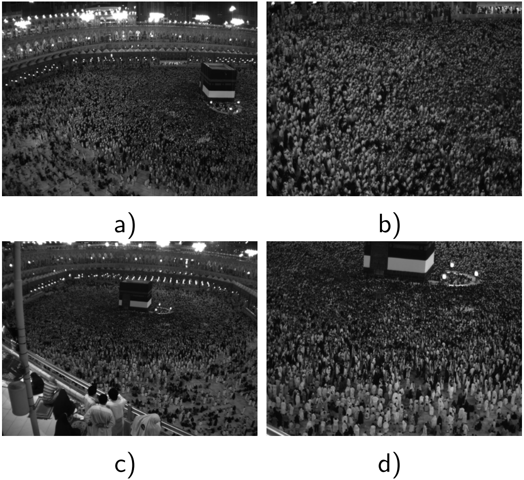
A set of images used for pose estimation in a simple indoor environment; the images in a) correspond to C_1^l and C_2^l, and the images in b) show the images captured by C_1^s and C_2^s. Both pairs of images have been matched using SIFT; the first set of matches are necessary for the algorithm, whilst the second set is used *exclusively as ground truth for validating the result of the algorithm*.

A small scale scenario

Table: Relative poses between C_1^s and C_2^s. The Euler angles are expressed in degrees, and the mean reprojection errors in pixels. Tilde values represent estimations prior to the BA procedure, and hat values denote estimations refined by BA. The difference between the two rows consists in the initialization of BA; in the first case we use the SIFT matches depicted in Figure 1b), whilst in the second case we use the result of our algorithm.

$(\tilde{\psi}; \tilde{\theta}; \tilde{\phi})$	$\tilde{\mathbf{C}}$	$\tilde{\epsilon}$	$(\hat{\psi}; \hat{\theta}; \hat{\phi})$	$\hat{\mathbf{C}}$	$\hat{\epsilon}$	**Iter.**	**Observation**
$\begin{pmatrix} 24.13 \\ 21.04 \\ 10.67 \end{pmatrix}$	$\begin{pmatrix} -0.89 \\ -0.30 \\ 0.33 \end{pmatrix}$	37.16	$\begin{pmatrix} 23.95 \\ 21.13 \\ 3.74 \end{pmatrix}$	$\begin{pmatrix} -0.79 \\ -0.25 \\ 0.55 \end{pmatrix}$	0.199	37	Base solution
$\begin{pmatrix} 23.85 \\ 16.42 \\ 3.53 \end{pmatrix}$	$\begin{pmatrix} -0.77 \\ -0.23 \\ 0.59 \end{pmatrix}$	0.489	$\begin{pmatrix} 23.95 \\ 21.13 \\ 3.74 \end{pmatrix}$	$\begin{pmatrix} -0.79 \\ -0.25 \\ 0.55 \end{pmatrix}$	0.199	25	Initialisation by $\mathbf{E}_{21}^s$

A large scale scene



a)

b)

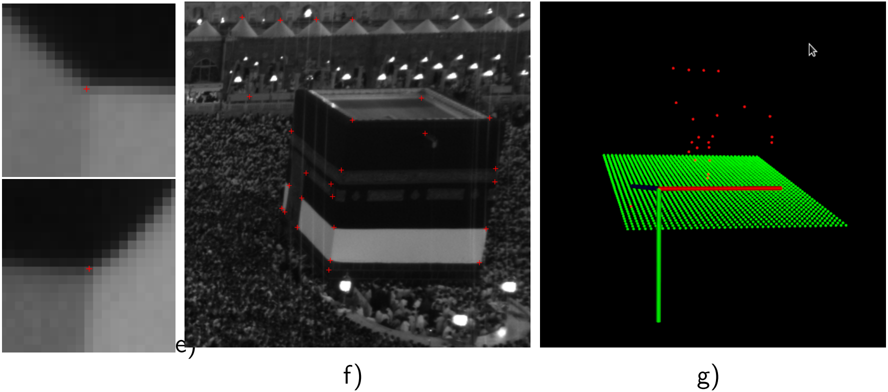
c)

d)

A set of images used for pose estimation; the images in a) and b) correspond to C_1^l and C_1^s, and c) and d) correspond to C_2^l and C_2^s respectively.

A large scale scene

Problem: FOV of change goes beyond the capabilities of SIFT, ASIFT etc.



e)

f)

g)

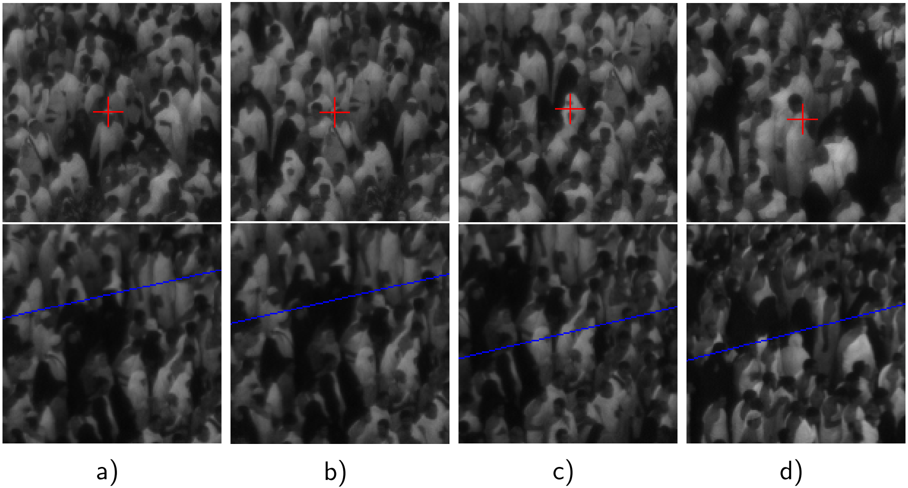
An example of user specified correspondences is illustrated in e). In f) we present the interest points used in the central region of one of the images, and in g) the inferred camera orientation (RGB axis for XYZ), with the approximate ground plane highlighted in green, for easier visualization.

A large scale scene

Table: Relative poses between analysis cameras placed on different rigs (first row), and between cameras placed on the same rig (rows 2-5). Tilde values represent estimations prior to the BA procedure, and hat values denote estimations refined by BA. The difference between the rows 2-3 and 4-5 consists in the initialization of the BA; in the first case we use the stereo calibration, whilst in the second case we use directly the images, in the same way as for the first row initialization.

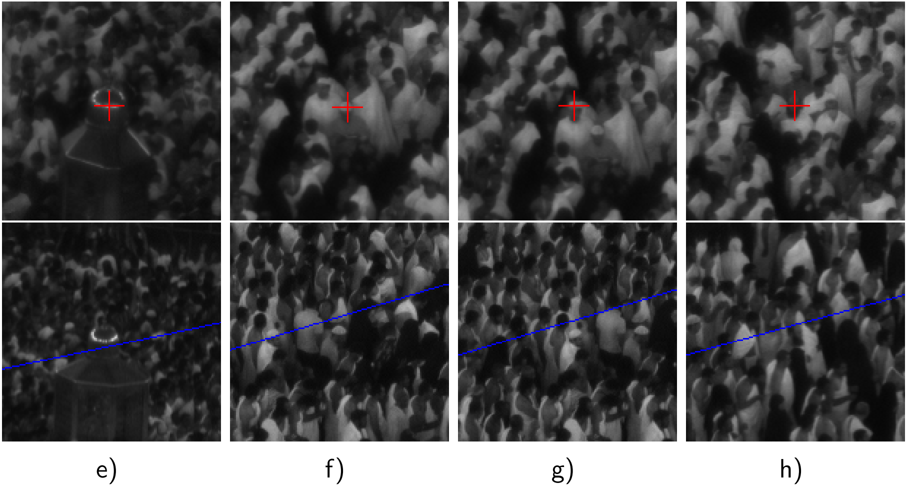
Cam. pair	$(\tilde{\boldsymbol{\psi}};\tilde{\boldsymbol{\theta}};\tilde{\boldsymbol{\phi}})$	$\tilde{\mathbf{C}}$	$\tilde{\boldsymbol{\epsilon}}$	$(\hat{\boldsymbol{\psi}};\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\phi}})$	$\hat{\mathbf{C}}$	$\hat{\boldsymbol{\epsilon}}$	**Observations**
$C_1^l \Rightarrow C_2^l$	$\begin{pmatrix}53.27\\71.52\\32.94\end{pmatrix}$	$\begin{pmatrix}-0.78\\-0.19\\0.59\end{pmatrix}$	4.00	$\begin{pmatrix}59.68\\69.11\\42.75\end{pmatrix}$	$\begin{pmatrix}-0.81\\-0.18\\0.55\end{pmatrix}$	0.25	Manual Init.
$C_1^l \Rightarrow C_1^s$	$\begin{pmatrix}-0.37\\-0.58\\0.51\end{pmatrix}$	$\begin{pmatrix}0.79\\0.00\\-0.62\end{pmatrix}$	1.017	$\begin{pmatrix}-0.33\\-0.34\\0.48\end{pmatrix}$	$\begin{pmatrix}0.11\\-0.01\\-0.99\end{pmatrix}$	0.076	Stereo Calib. Init.
$C_2^l \Rightarrow C_2^s$	$\begin{pmatrix}-0.19\\0.72\\0.23\end{pmatrix}$	$\begin{pmatrix}0.94\\0.05\\-0.34\end{pmatrix}$	1.661	$\begin{pmatrix}-0.09\\0.71\\0.12\end{pmatrix}$	$\begin{pmatrix}0.57\\0.23\\0.78\end{pmatrix}$	0.252	Stereo Calib. Init.
$C_1^l \Rightarrow C_1^s$	$\begin{pmatrix}-0.36\\-0.23\\0.43\end{pmatrix}$	$\begin{pmatrix}0.01\\-0.06\\0.99\end{pmatrix}$	0.096	$\begin{pmatrix}-0.33\\-0.28\\0.47\end{pmatrix}$	$\begin{pmatrix}0.06\\-0.02\\-0.99\end{pmatrix}$	0.084	SIFT Matching
$C_2^l \Rightarrow C_2^s$	$\begin{pmatrix}-0.16\\0.74\\0.10\end{pmatrix}$	$\begin{pmatrix}-0.02\\-0.01\\-0.99\end{pmatrix}$	0.097	$\begin{pmatrix}-0.13\\0.75\\0.10\end{pmatrix}$	$\begin{pmatrix}-0.01\\0.00\\-0.99\end{pmatrix}$	0.087	SIFT Matching

Epipolar projection



A number of pixel-epipolar line correspondences between the two analysis cameras presented in Figure 3b) and 3d). Ideally, the correspondent of a point highlighted by the red cross in the upper row should be situated along the blue epipolar line visible in the lower row image.

Epipolar projection



A number of pixel-epipolar line correspondences between the two analysis cameras presented in Figure 3b) and 3d). Ideally, the correspondent of a point highlighted by the red cross in the upper row should be situated along the blue epipolar line visible in the lower row image.

Outline

1. Context
2. The proposed algorithm
3. Experimental results
4. Conclusions and future work

Conclusions and future work

Conclusions

- A new method for aligning multiple cameras analysing a homogeneous scene
- Settings where for practical reasons calibration pattern/object based registrations are not possible
- Avoid making any assumptions about the homogeneous region we analyse
- Validation in an indoor environment and in a large scale scenario
- Particularly adapted to multiple camera tracking and other data fusion strategies

Perspectives

- Multiple camera based map fusion and tracking
- Online re-estimation (robustness to camera shaking)
- Exploit the pyramidal information in order to estimate the impact on image processing algorithms