Spatio-temporal Consistency for Head Detection in High-Density Scenes

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

[1]Autonomous Systems Group, Université Paris Sud, France
[2]Università di Pavia, Italy
[3]Communications and Signal Processing Group, Dept. of Electrical and Electronic Engineering, Imperial College London, UK

Outline

Outline

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)
- However, particle detection in single views is an essential step, and can not be circumvented

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)
- However, particle detection in single views is an essential step, and can not be circumvented

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)
- However, particle detection in single views is an essential step, and can not be circumvented

The context of this work

Modelling high-density crowded scenes

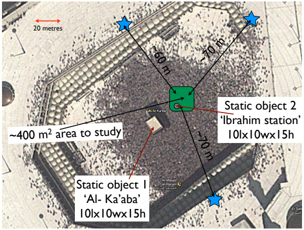
- Understanding pedestrian dynamics at high densities
- Understanding how instabilities may build up
- Micro-analysis: in order to model the system, the particles (pedestrians) must be tracked individually

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)
- However, particle detection in single views is an essential step, and can not be circumvented

An ideal testing scenario

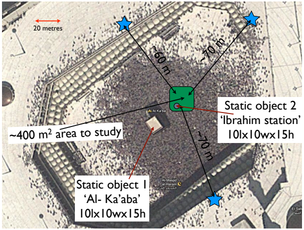
Very challenging but also suitable setting: Makkah



- ✓ Major interest for improving security
- ✓ Constant people flow
- ✗ High security, important logistical constraints
- ✗ Very large scale scene

An ideal testing scenario

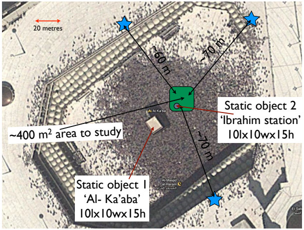
Very challenging but also suitable setting: Makkah



- ✓ Major interest for improving security
- ✓ Constant people flow
- ✗ High security, important logistical constraints
- ✗ Very large scale scene

An ideal testing scenario

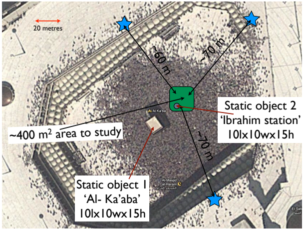
Very challenging but also suitable setting: Makkah



- ✓ Major interest for improving security
- ✓ Constant people flow
- ✗ High security, important logistical constraints
- ✗ Very large scale scene

An ideal testing scenario

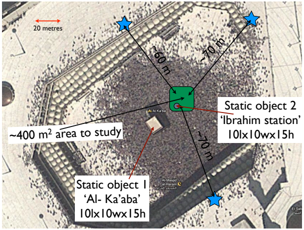
Very challenging but also suitable setting: Makkah



- ✓ Major interest for improving security
- ✓ Constant people flow
- ✗ High security, important logistical constraints
- ✗ Very large scale scene

An ideal testing scenario

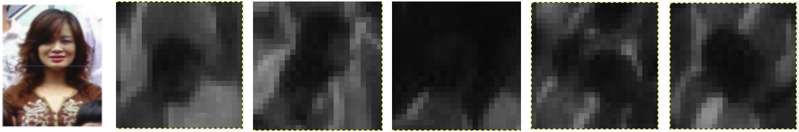
Very challenging but also suitable setting: Makkah



- ✓ Major interest for improving security
- ✓ Constant people flow
- ✗ High security, important logistical constraints
- ✗ Very large scale scene

The standard strategy

Discriminative learning: used extensively for pedestrian detection in uncongested and moderately crowded contexts



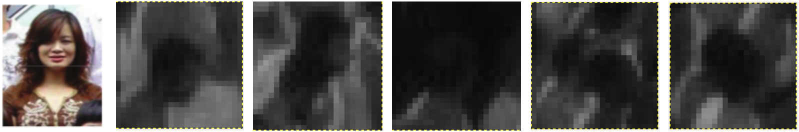
(a) (b) (c) (d) (e) (f)

- In 1(a): for comparison, an image used[1] for learning the head-shoulder shape; typical patch sizes in the literature: 32×32 to 48×64
- In 1(b): a typical well contrasted head in our dataset; significantly lower resolution per target
- In 1(c) - 1(f): low contrast between close targets, between targets and the dynamic background, strong occlusions

[1] Li et al.: Head-shoulder based gender recognition. ICIP 2013

The standard strategy

Discriminative learning: used extensively for pedestrian detection in uncongested and moderately crowded contexts



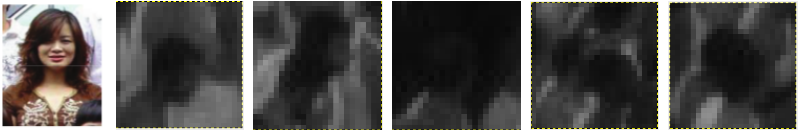
(a) (b) (c) (d) (e) (f)

- In 1(a): for comparison, an image used[1] for learning the head-shoulder shape; typical patch sizes in the literature: 32×32 to 48×64
- In 1(b): a typical well contrasted head in our dataset; significantly lower resolution per target
- In 1(c) - 1(f): low contrast between close targets, between targets and the dynamic background, strong occlusions

[1] Li et al.: Head-shoulder based gender recognition. ICIP 2013

The standard strategy

Discriminative learning: used extensively for pedestrian detection in uncongested and moderately crowded contexts



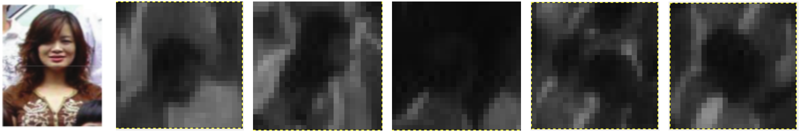
(a) (b) (c) (d) (e) (f)

- In 1(a): for comparison, an image used[1] for learning the head-shoulder shape; typical patch sizes in the literature: 32×32 to 48×64
- In 1(b): a typical well contrasted head in our dataset; significantly lower resolution per target
- In 1(c) - 1(f): low contrast between close targets, between targets and the dynamic background, strong occlusions

[1] Li et al.: Head-shoulder based gender recognition. ICIP 2013

The standard strategy

Discriminative learning: used extensively for pedestrian detection in uncongested and moderately crowded contexts

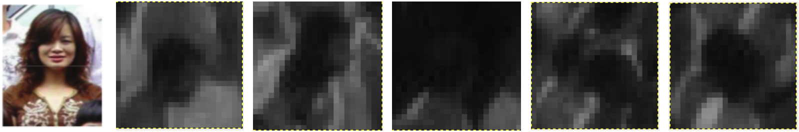


(a) (b) (c) (d) (e) (f)

- Close to the limit of interpretation of a human subject
- Can we still apply discriminative learning and obtain something meaningful in these extreme settings?
- Or should we fundamentally change the way we approach this problem?

The standard strategy

Discriminative learning: used extensively for pedestrian detection in uncongested and moderately crowded contexts

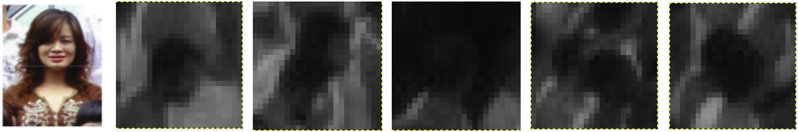


(a) (b) (c) (d) (e) (f)

- Close to the limit of interpretation of a human subject
- Can we still apply discriminative learning and obtain something meaningful in these extreme settings?
- Or should we fundamentally change the way we approach this problem?

The standard strategy

Discriminative learning: used extensively for pedestrian detection in uncongested and moderately crowded contexts



(a) (b) (c) (d) (e) (f)

- Close to the limit of interpretation of a human subject
- Can we still apply discriminative learning and obtain something meaningful in these extreme settings?
- Or should we fundamentally change the way we approach this problem?

Outline

Discriminative learning

The descriptor

- We rely on the HOG (among 2-3 other “classical” alternatives)
- Two main assumptions:
  - Size of the targets: a disk of a three-four pixel average radius
  - Occlusions are frequent and strong
- Significant impact of the window size parameter

Discriminative learning

The descriptor

- We rely on the HOG (among 2-3 other “classical” alternatives)
- Two main assumptions:
  - Size of the targets: a disk of a three-four pixel average radius
  - Occlusions are frequent and strong
- Significant impact of the window size parameter

Discriminative learning

The descriptor

- We rely on the HOG (among 2-3 other “classical” alternatives)
- Two main assumptions:
  - Size of the targets: a disk of a three-four pixel average radius
  - Occlusions are frequent and strong
- Significant impact of the window size parameter

Discriminative learning

The descriptor

- We rely on the HOG (among 2-3 other “classical” alternatives)
- Two main assumptions:
  - Size of the targets: a disk of a three-four pixel average radius
  - Occlusions are frequent and strong
- Significant impact of the window size parameter

Discriminative learning

The descriptor

- We rely on the HOG (among 2-3 other “classical” alternatives)
- Two main assumptions:
  - Size of the targets: a disk of a three-four pixel average radius
  - Occlusions are frequent and strong
- Significant impact of the window size parameter

The learning task

- We rely on an SVM classifier, and we consider two different kernels:
  - A linear classifier $K_L(h_1, h_2) = \langle h_1, h_2 \rangle$
  - The Histogram Intersection Kernel (HIK) function
- Pixel-wise classification and transfer of the binary classifier decision into a probability estimation $p_{i,j}$

Discriminative learning

The descriptor

- We rely on the HOG (among 2-3 other “classical” alternatives)
- Two main assumptions:
  - Size of the targets: a disk of a three-four pixel average radius
  - Occlusions are frequent and strong
- Significant impact of the window size parameter

The learning task

- We rely on an SVM classifier, and we consider two different kernels:
  - A linear classifier $K_L(h_1, h_2) = \langle h_1, h_2 \rangle$
  - The Histogram Intersection Kernel (HIK) function
- Pixel-wise classification and transfer of the binary classifier decision into a probability estimation $p_{i,j}$

Discriminative learning

The descriptor

- We rely on the HOG (among 2-3 other "classical" alternatives)
- Two main assumptions:
  - Size of the targets: a disk of a three-four pixel average radius
  - Occlusions are frequent and strong
- Significant impact of the window size parameter

The learning task

- We rely on an SVM classifier, and we consider two different kernels:
  - A linear classifier $K_L(h_1, h_2) = \langle h_1, h_2 \rangle$
  - The Histogram Intersection Kernel (HIK) function

$$K_I(h_1, h_2) = \sum_{i=1}^{dim} \min[h_1(i), h_2(i)]$$

Discriminative learning

The descriptor

- We rely on the HOG (among 2-3 other "classical" alternatives)
- Two main assumptions:
  - Size of the targets: a disk of a three-four pixel average radius
  - Occlusions are frequent and strong
- Significant impact of the window size parameter

The learning task

- We rely on an SVM classifier, and we consider two different kernels:
  - A linear classifier $K_L(h_1, h_2) = \langle h_1, h_2 \rangle$
  - The Histogram Intersection Kernel (HIK) function
- Pixel-wise classification and transfer of the binary classifier decision into a probability estimation $p_{i,j}$

Evaluation criterion

- Getting pixel level ground truth is very costly
- A human user clicks exhaustively and as accurately as possible in the centre of the targets
- We expect pixels located in discretized disks of radius r around ground truth points be classified as positives
  - true positives: $p_i \wedge p_j$
  - false positives: $p_i \wedge \neg p_j$
  - false negatives: $\neg p_i \wedge p_j$
  - true negatives: $\neg (p_i \vee p_j)$

Evaluation criterion

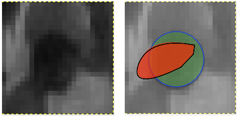
- Getting pixel level ground truth is very costly
- A human user clicks exhaustively and as accurately as possible in the centre of the targets
- We expect pixels located in discretized disks of radius r around ground truth points be classified as positives
  - true positives: $p_i \wedge p_j$
  - false positives: $p_i \wedge \neg p_j$
  - false negatives: $\neg p_i \wedge p_j$
  - true negatives: $\neg (p_i \vee p_j)$

Evaluation criterion

- Getting pixel level ground truth is very costly
- A human user clicks exhaustively and as accurately as possible in the centre of the targets
- We expect pixels located in discretized disks of radius r around ground truth points be classified as positives
  - true positives: $p_i \wedge p_j$
  - false positives: $p_i \wedge \neg\, p_j$
  - false negatives: $\neg\, p_i \wedge p_j$
  - true negatives: $\neg\, (p_i \vee p_j)$

Evaluation criterion

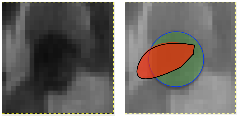
- Getting pixel level ground truth is very costly
- A human user clicks exhaustively and as accurately as possible in the centre of the targets
- We expect pixels located in discretized disks of radius r around ground truth points be classified as positives



- p_i: classified as positive; p_j: labelled as positive
  - true positives: $p_i \wedge p_j$
  - false positives: $p_i \wedge \neg p_j$
  - false negatives: $\neg p_i \wedge p_j$
  - true negatives: $\neg (p_i \vee p_j)$

Evaluation criterion

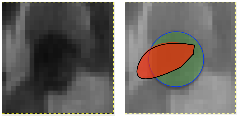
- Getting pixel level ground truth is very costly
- A human user clicks exhaustively and as accurately as possible in the centre of the targets
- We expect pixels located in discretized disks of radius r around ground truth points be classified as positives



- p_i: classified as positive; p_j: labelled as positive
  - true positives: $p_i \wedge p_j$
  - false positives: $p_i \wedge \neg p_j$
  - false negatives: $\neg p_i \wedge p_j$
  - true negatives: $\neg (p_i \vee p_j)$

Evaluation criterion

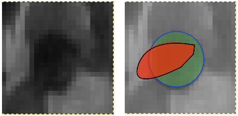
- Getting pixel level ground truth is very costly
- A human user clicks exhaustively and as accurately as possible in the centre of the targets
- We expect pixels located in discretized disks of radius r around ground truth points be classified as positives



- p_i: classified as positive; p_j: labelled as positive
  - true positives: $p_i \wedge p_j$
  - false positives: $p_i \wedge \neg p_j$
  - false negatives: $\neg p_i \wedge p_j$
  - true negatives: $\neg (p_i \vee p_j)$

Evaluation criterion

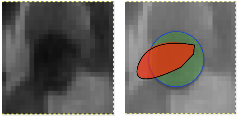
- Getting pixel level ground truth is very costly
- A human user clicks exhaustively and as accurately as possible in the centre of the targets
- We expect pixels located in discretized disks of radius r around ground truth points be classified as positives



- p_i: classified as positive; p_j: labelled as positive
  - true positives: $p_i \wedge p_j$
  - false positives: $p_i \wedge \neg p_j$
  - false negatives: $\neg p_i \wedge p_j$
  - true negatives: $\neg (p_i \vee p_j)$

Evaluation criterion

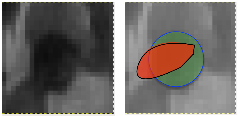
- Getting pixel level ground truth is very costly
- A human user clicks exhaustively and as accurately as possible in the centre of the targets
- We expect pixels located in discretized disks of radius r around ground truth points be classified as positives



- p_i: classified as positive; p_j: labelled as positive
  - true positives: $p_i \wedge p_j$
  - false positives: $p_i \wedge \neg p_j$
  - false negatives: $\neg p_i \wedge p_j$
  - true negatives: $\neg (p_i \vee p_j)$

Evaluation criterion

- Getting pixel level ground truth is very costly
- A human user clicks exhaustively and as accurately as possible in the centre of the targets
- We expect pixels located in discretized disks of radius r around ground truth points be classified as positives



- p_i: classified as positive; p_j: labelled as positive
  - true positives: $p_i \wedge p_j$
  - false positives: $p_i \wedge \neg p_j$
  - false negatives: $\neg p_i \wedge p_j$
  - true negatives: $\neg (p_i \vee p_j)$

Outline

Spatio-temporal consistency

Motivation

- Difficulty: the descriptor varies on a high-dimensional feature space

✗ The probability of the location may occasionally vary significantly

✓ We have video sequences, dynamics are low

✓ The detector response should be blob-like

Spatio-temporal consistency

Motivation

- Difficulty: the descriptor varies on a high-dimensional feature space

✗ The probability of the location may occasionally vary significantly

✓ We have video sequences, dynamics are low

✓ The detector response should be blob-like

Spatio-temporal consistency

Motivation

- Difficulty: the descriptor varies on a high-dimensional feature space
- ✗ The probability of the location may occasionally vary significantly
- ✓ We have video sequences, dynamics are low
- ✓ The detector response should be blob-like

Spatio-temporal consistency

Motivation

- Difficulty: the descriptor varies on a high-dimensional feature space
- ✗ The probability of the location may occasionally vary significantly
- ✓ We have video sequences, dynamics are low
- ✓ The detector response should be blob-like

Spatio-temporal consistency

Motivation

- Difficulty: the descriptor varies on a high-dimensional feature space
- ✗ The probability of the location may occasionally vary significantly
- ✓ We have video sequences, dynamics are low
- ✓ The detector response should be blob-like

Temporal consistency

- Assumption: short-term variations in the probability values should be small for pixels belonging genuinely to targets
- Secondarily, a target consists in multiple connected pixels, so encourage clustered responses in the probability distribution
- Not a tracking algorithm: avoid inference at object level, and we provide a reliable *pixel-wise* label for head detection

Spatio-temporal consistency

Motivation

- Difficulty: the descriptor varies on a high-dimensional feature space
- ✗ The probability of the location may occasionally vary significantly
- ✓ We have video sequences, dynamics are low
- ✓ The detector response should be blob-like

Temporal consistency

- Assumption: short-term variations in the probability values should be small for pixels belonging genuinely to targets
- Secondarily, a target consists in multiple connected pixels, so encourage clustered responses in the probability distribution
- Not a tracking algorithm: avoid inference at object level, and we provide a reliable *pixel-wise* label for head detection

Spatio-temporal consistency

Motivation

- Difficulty: the descriptor varies on a high-dimensional feature space
- ✗ The probability of the location may occasionally vary significantly
- ✓ We have video sequences, dynamics are low
- ✓ The detector response should be blob-like

Temporal consistency

- Assumption: short-term variations in the probability values should be small for pixels belonging genuinely to targets
- Secondarily, a target consists in multiple connected pixels, so encourage clustered responses in the probability distribution
- Not a tracking algorithm: avoid inference at object level, and we provide a reliable *pixel-wise* label for head detection

Spatio-temporal consistency

Motivation

- Difficulty: the descriptor varies on a high-dimensional feature space
- ✗ The probability of the location may occasionally vary significantly
- ✓ We have video sequences, dynamics are low
- ✓ The detector response should be blob-like

Temporal consistency

- Assumption: short-term variations in the probability values should be small for pixels belonging genuinely to targets
- Secondarily, a target consists in multiple connected pixels, so encourage clustered responses in the probability distribution
- Not a tracking algorithm: avoid inference at object level, and we provide a reliable *pixel-wise* label for head detection

Temporal consistency

Steps

- Small movement $\Rightarrow$ reliable dense optical flow
- Consider a detection threshold τ, a pixel $l^t_{i,j}$, and corresponding projections $l^{t+k}_{i,j}$, where $-N \leq k \leq N$
  1. Spatial regularization in the immediate neighbourhood of the projections: maximal voting (8-adjacency) to get $l^{t+k}_{i,j}$ of $l^{t+k}_{i,j}$
  2. maximal vote on the set

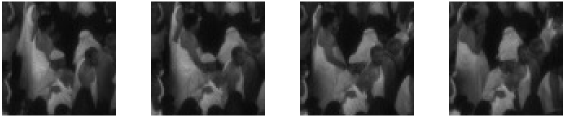
$$L^t_{i,j} = \{l^{t+k}_{i,j}\}_{-N \leq k \leq N}$$

and assign the result to $l^t_{i,j}$

Temporal consistency

Steps

- Small movement $\Rightarrow$ reliable dense optical flow



- Consider a detection threshold τ, a pixel $I^{t}_{i,j}$, and corresponding projections $I^{t+k}_{i,j}$, where $-N \leq k \leq N$
  1. Spatial regularization in the immediate neighbourhood of the projections: maximal voting (8-adjacency) to get $I^{t+k}_{i,j}$ of $I^{t+k}_{i,j}$
  2. maximal vote on the set

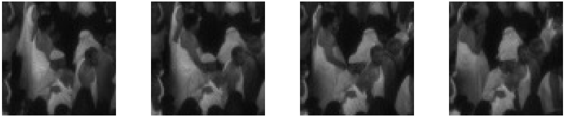
$$L^{t}_{i,j} = \{I^{t+k}_{i,j}\}_{-N \leq k \leq N}$$

and assign the result to $I^{t}_{i,j}$

Temporal consistency

Steps

- Small movement $\Rightarrow$ reliable dense optical flow



- Consider a detection threshold τ, a pixel $I^{t}_{i,j}$, and corresponding projections $I^{t+k}_{i,j}$, where $-N \leq k \leq N$
  1. Spatial regularization in the immediate neighbourhood of the projections: maximal voting (8-adjacency) to get $I^{t+k}_{i,j}$ of $I^{t+k}_{i,j}$
  2. maximal vote on the set

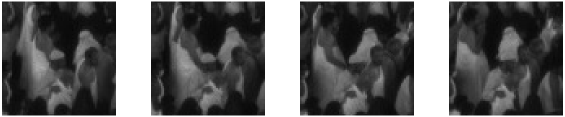
$$L^{t}_{i,j} = \{I^{t+k}_{i,j}\}_{-N \leq k \leq N}$$

and assign the result to $I^{t}_{i,j}$

Temporal consistency

Steps

- Small movement $\Rightarrow$ reliable dense optical flow



- Consider a detection threshold τ, a pixel $I_{i,j}^{t}$, and corresponding projections $I_{i,j}^{t+k}$, where $-N \leq k \leq N$
  1. Spatial regularization in the immediate neighbourhood of the projections: maximal voting (8-adjacency) to get $l_{i,j}^{t+k}$ of $I_{i,j}^{t+k}$
  2. maximal vote on the set

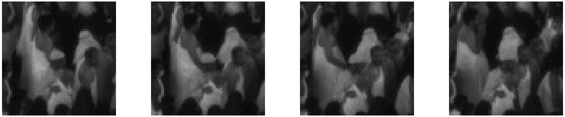
$$L_{i,j}^{t} = \{l_{i,j}^{t+k}\}_{-N \leq k \leq N}$$

and assign the result to $I_{i,j}^{t}$

Temporal consistency

Steps

- Small movement $\Rightarrow$ reliable dense optical flow



- Consider a detection threshold τ, a pixel $l^t_{i,j}$, and corresponding projections $l^{t+k}_{i,j}$, where $-N \leq k \leq N$
  1. Spatial regularization in the immediate neighbourhood of the projections: maximal voting (8-adjacency) to get $l^{t+k}_{i,j}$ of $l^{t+k}_{i,j}$
  2. maximal vote on the set

$$L^t_{i,j} = \{l^{t+k}_{i,j}\}_{-N \leq k \leq N}$$

and assign the result to $l^t_{i,j}$

Spatial consistency

Steps

- We refine a posteriori the pixel classification $l_{i,j}^t$
- We assume a Markov random field (MRF) over the pixel states.
- We consider a basic symmetric neighborhood structure based on 4-adjacency

$$N_{i,j}^t = \{l_{i-1,j}^t, l_{i+1,j}^t, l_{i,j-1}^t, l_{i,j+1}^t\}$$

- We consider as observation set the current probability map associating to the pixel $l_{i,j}^t$ the values $p_{i,j}^t \in [0,1]$ provided by the classifier

Spatial consistency

Steps

- We refine a posteriori the pixel classification $l_{i,j}^t$
- We assume a Markov random field (MRF) over the pixel states.
- We consider a basic symmetric neighborhood structure based on 4-adjacency

$$N_{i,j}^t = \{l_{i-1,j}^t, l_{i+1,j}^t, l_{i,j-1}^t, l_{i,j+1}^t\}$$

- We consider as observation set the current probability map associating to the pixel $l_{i,j}^t$ the values $p_{i,j}^t \in [0,1]$ provided by the classifier

Spatial consistency

Steps

- We refine a posteriori the pixel classification $l^t_{i,j}$
- We assume a Markov random field (MRF) over the pixel states.
- We consider a basic symmetric neighborhood structure based on 4-adjacency

$$N^t_{i,j} = \{l^t_{i-1,j}, l^t_{i+1,j}, l^t_{i,j-1}, l^t_{i,j+1}\}$$

- We consider as observation set the current probability map associating to the pixel $l^t_{i,j}$ the values $p^t_{i,j} \in [0, 1]$ provided by the classifier

Spatial consistency

Steps

- We refine a posteriori the pixel classification $l^t_{i,j}$
- We assume a Markov random field (MRF) over the pixel states.
- We consider a basic symmetric neighborhood structure based on 4-adjacency

$$N^t_{i,j} = \{l^t_{i-1,j}, l^t_{i+1,j}, l^t_{i,j-1}, l^t_{i,j+1}\}$$

- We consider as observation set the current probability map associating to the pixel $l^t_{i,j}$ the values $p^t_{i,j} \in [0,1]$ provided by the classifier

Outline

Experimental setup

Parameters

- High-density images acquired at Makkah
- Training: 1032 patches used as positive and negative examples
- Descriptor window size was set to 24×24
- Training with the linear kernel: 241 support vectors
- Training with the HI kernel: 343 support vectors

Experimental setup

Parameters

- High-density images acquired at Makkah
- Training: 1032 patches used as positive and negative examples
- Descriptor window size was set to 24×24
- Training with the linear kernel: 241 support vectors
- Training with the HI kernel: 343 support vectors

Experimental setup

Parameters

- High-density images acquired at Makkah
- Training: 1032 patches used as positive and negative examples
- Descriptor window size was set to 24×24
- Training with the linear kernel: 241 support vectors
- Training with the HI kernel: 343 support vectors

Experimental setup

Parameters

- High-density images acquired at Makkah
- Training: 1032 patches used as positive and negative examples
- Descriptor window size was set to 24×24
- Training with the linear kernel: 241 support vectors
- Training with the HI kernel: 343 support vectors

Experimental setup

Parameters

- High-density images acquired at Makkah
- Training: 1032 patches used as positive and negative examples
- Descriptor window size was set to 24×24
- Training with the linear kernel: 241 support vectors
- Training with the HI kernel: 343 support vectors

Experimental setup

Parameters

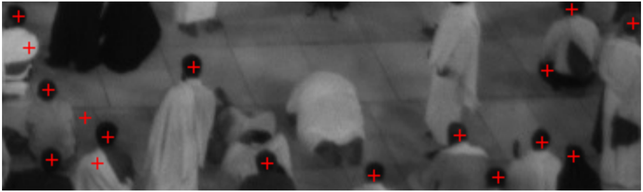
- High-density images acquired at Makkah
- Training: 1032 patches used as positive and negative examples
- Descriptor window size was set to 24×24
- Training with the linear kernel: 241 support vectors
- Training with the HI kernel: 343 support vectors

The cluttered context has a significant impact on classifier performance. Procedure: detection probability map, thresholding and non-maximal suppression (linear kernel).

Experimental setup

Parameters

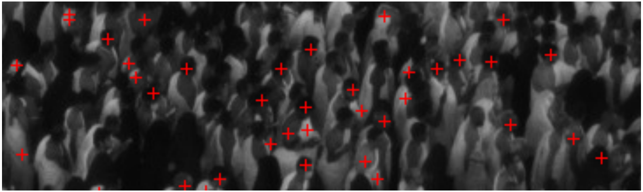
- High-density images acquired at Makkah
- Training: 1032 patches used as positive and negative examples
- Descriptor window size was set to 24×24
- Training with the linear kernel: 241 support vectors
- Training with the HI kernel: 343 support vectors



Experimental setup

Parameters

- High-density images acquired at Makkah
- Training: 1032 patches used as positive and negative examples
- Descriptor window size was set to 24×24
- Training with the linear kernel: 241 support vectors
- Training with the HI kernel: 343 support vectors



Experimental setup

Parameters

- High-density images acquired at Makkah
- Training: 1032 patches used as positive and negative examples
- Descriptor window size was set to 24×24
- Training with the linear kernel: 241 support vectors
- Training with the HI kernel: 343 support vectors

- What threshold should we use?
- Do we need to threshold for our aims?
- What about regularization?
- Our suggestion: postpone as much as possible in the decision process the steps that lead to information loss (thresholding, non-maximal suppression)

Experimental setup

Parameters

- High-density images acquired at Makkah
- Training: 1032 patches used as positive and negative examples
- Descriptor window size was set to 24×24
- Training with the linear kernel: 241 support vectors
- Training with the HI kernel: 343 support vectors

- What threshold should we use?
- Do we need to threshold for our aims?
- What about regularization?
- Our suggestion: postpone as much as possible in the decision process the steps that lead to information loss (thresholding, non-maximal suppression)

Experimental setup

Parameters

- High-density images acquired at Makkah
- Training: 1032 patches used as positive and negative examples
- Descriptor window size was set to 24×24
- Training with the linear kernel: 241 support vectors
- Training with the HI kernel: 343 support vectors

- What threshold should we use?
- Do we need to threshold for our aims?
- What about regularization?
- Our suggestion: postpone as much as possible in the decision process the steps that lead to information loss (thresholding, non-maximal suppression)

Experimental setup

Parameters

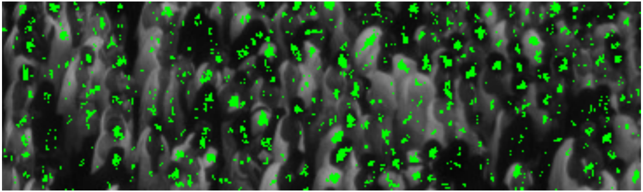
- High-density images acquired at Makkah
- Training: 1032 patches used as positive and negative examples
- Descriptor window size was set to 24×24
- Training with the linear kernel: 241 support vectors
- Training with the HI kernel: 343 support vectors

- What threshold should we use?
- Do we need to threshold for our aims?
- What about regularization?
- Our suggestion: postpone as much as possible in the decision process the steps that lead to information loss (thresholding, non-maximal suppression)

Experimental setup

Parameters

- High-density images acquired at Makkah
- Training: 1032 patches used as positive and negative examples
- Descriptor window size was set to 24×24
- Training with the linear kernel: 241 support vectors
- Training with the HI kernel: 343 support vectors



ROC analysis of the detector

Procedure

- We define a ground truth set consisting of 132 particles
- We consider different ground truth radii $0 \leq r \leq 4$
- The threshold τ is mapped over $\tau \in [0, 1]$

ROC analysis of the detector

Procedure

- We define a ground truth set consisting of 132 particles
- We consider different ground truth radii $0 \leq r \leq 4$
- The threshold τ is mapped over $\tau \in [0, 1]$

ROC analysis of the detector

Procedure

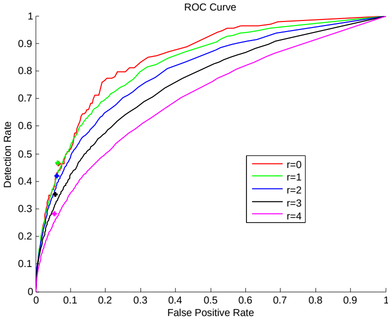
- We define a ground truth set consisting of 132 particles
- We consider different ground truth radii $0 \leq r \leq 4$
- The threshold τ is mapped over $\tau \in [0, 1]$

ROC analysis of the detector

Procedure

- We define a ground truth set consisting of 132 particles
- We consider different ground truth radii $0 \leq r \leq 4$
- The threshold τ is mapped over $\tau \in [0, 1]$

Linear kernel:

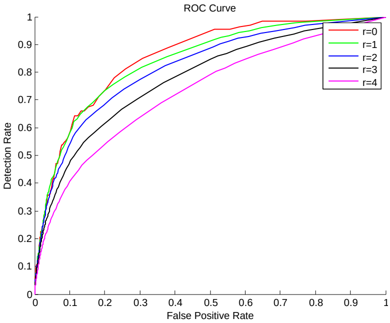


ROC analysis of the detector

Procedure

- We define a ground truth set consisting of 132 particles
- We consider different ground truth radii $0 \leq r \leq 4$
- The threshold τ is mapped over $\tau \in [0, 1]$

Linear kernel + temporal consistency check:

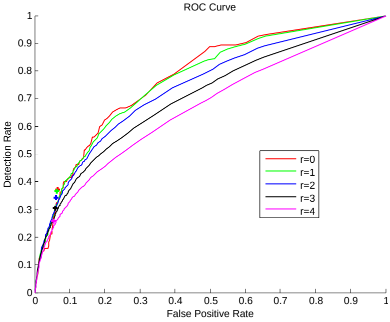


ROC analysis of the detector

Procedure

- We define a ground truth set consisting of 132 particles
- We consider different ground truth radii $0 \leq r \leq 4$
- The threshold τ is mapped over $\tau \in [0, 1]$

HI kernel:

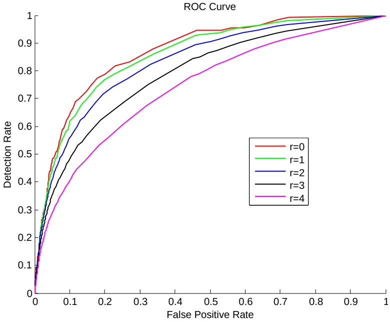


ROC analysis of the detector

Procedure

- We define a ground truth set consisting of 132 particles
- We consider different ground truth radii $0 \leq r \leq 4$
- The threshold τ is mapped over $\tau \in [0, 1]$

HI kernel + temporal consistency check:



Outline

Conclusions and future work

Conclusions

- Discriminative learning may be employed, even in extremely cluttered environments, to provide target cues to tracking algorithms
- HIK + temporal information provide the most effective results
- ROC curves highlight the trade-off between the risk of target miss and the presence of false positives
- Consistent detection probability maps which present a plateau response in target locations
- Particularly adapted to multiple camera tracking and other data fusion strategies

Perspectives

- A formal framework for regularization with a reasonable computational cost
- The impact of the topology on the classifier performance
- Multiple camera based map fusion and tracking
- Difficulties related to datasets, validation and training