A deep learning strategy for wide-area surveillance

17/05/2016

Mr Alessandro Borgia

Supervisor: Prof Neil Robertson

Heriot-Watt University – EPS/ISSS – Visionlab

Roke Manor Research partnership

Outline

- The proposed re-identification system:
  - A boostrap process for tracking: unifying tracking and deep learning-based re-identifications
  - Intra-camera tracking scheme
  - Inter-camera tracking: time transition distributions estimation over the network
- Cross-Input Neighborhood Differences (CIND) CNN:
- A more flexible approach for CNN:
  - Going deeper by residual learning
  - Triplet network training scheme
  - Batch normalization
- Simulations
- Visualizing deep features
- References

- Outline
- Motivation
- Proposed system
- Intra-camera tracking
- Time transition distribution
- Spacial distribution estimation
- Advantages
- CIND-CNN
- CUHK-03 dataset
- A more flexible approach
  - Residual learning
  - Triplet network
  - Batch norm.
- Simulations
- Features appearance
- Next step

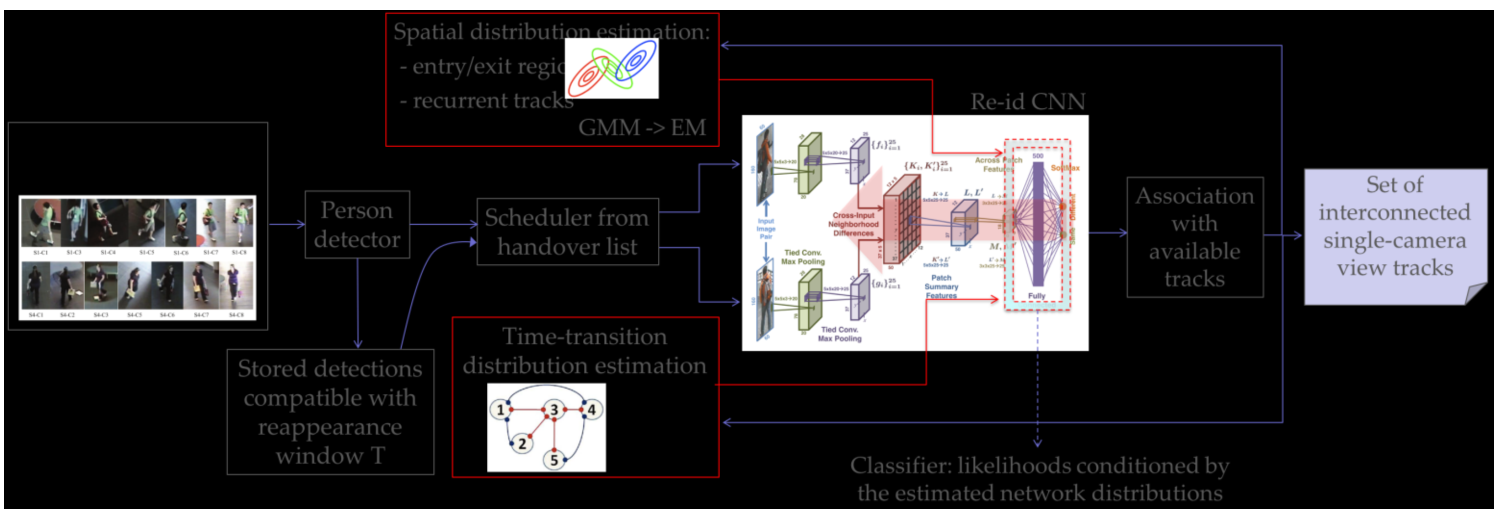
Motivation

- Outline
- Motivation
- Proposed system
- Intra-camera tracking
- Time transition distribution
- Spacial distribution estimation
- Advantages
- CIND-CNN
- CUHK-03 dataset
- A more flexible approach
  - Residual learning
  - Triplet network
  - Batch norm.
- Simulations
- Features appearance
- Next step

- **Context**: people tracking in multiple non-overlapping cameras
- **Problem**: dealing with targets disappearing for extended periods of time (long occlusions)
- **Challenges** arising in different camera views: complex variations of lightings, poses, viewpoints, occlusions.
- **Traditional approaches**: engineering hand-crafted features
- **Actual approach**: employing a deep learning-based (DL) re-identification strategy
- **Why?**: a deep architecture allows to model effectively the mixture of complex multimodal photometric and geometric transforms that targets undergo.
- **Novelty**: the proposed DL-based re-identification scheme is proposed as a boostrap process for the inter-camera tracking task, defining a unified framework

The proposed system

- Iterative adaptive interaction between the re-identification and tracking tasks
- Effect: boosting each other: more powerful tracking capabilities in presence of disappearing targets and
- The re-id stage feeds the process of automatic refinement of the logical topology and temporal interdependences of the network (automatically learned from observations)
- The temporal distributions, by feeding the CNN classifier (and back-tuning the weights accordingly) enable the CNN to take more reliable context-aware re-id decisions.



- Outline
- Motivation
- Proposed system
- Intra-camera tracking
- Time transition distribution
- Spacial distribution estimation
- Advantages
- CIND-CNN
- CUHK-03 dataset
- A more flexible approach
  - Residual learning
  - Triplet network
  - Batch norm.
- Simulations
- Features appearance
- Next step

Intra-camera tracking scheme

- Investigated context: a wide area surveillance network with unknown, unconstrained topology and non-calibrated static CCTV cameras
- Tracking based only on re-identifications by a CNN.
- Gathering entry and exit points of all the built trajectories
- Estimation of the entry/exit regions by Gaussian Mixture Model and Expectation Maximization algorithm
- Entry/exit points represent the network nodes according to which to buid the network logical topology

- Outline
- Motivation
- Proposed system
- Intra-camera tracking
- Time transition distribution
- Spacial distribution estimation
- Advantages
- CIND-CNN
- CUHK-03 dataset
- A more flexible approach
  - Residual learning
  - Triplet network
  - Batch norm.
- Simulations
- Features appearance
- Next step

Roke Part of the Chemring Group

Time transition distribution over all links

Time transition distribution over all links:

Network cameras: $C_1, \ldots, C_N$

$p_{ab}(t)$: probability of reappearance from Ca→Cb at time t

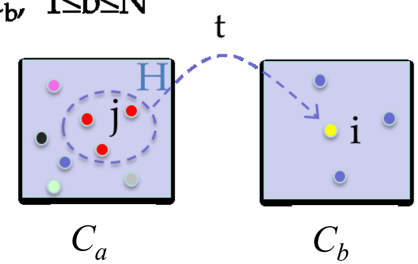
i: new target appeared at time t in view C_b, $1 \leq b \leq N$

j: target leaving view C_a in [t-T, t+T],

T: reappearance window

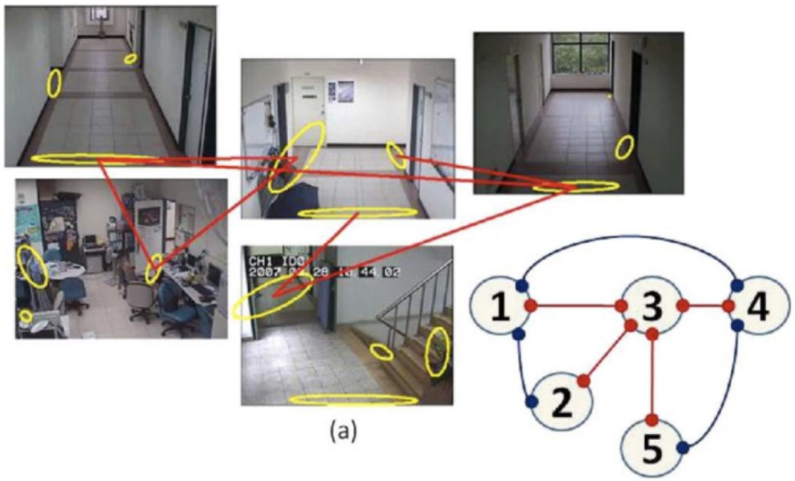
H: handover list (leaving targets)

id(n): identity of target n



$$p_{ab}(t) = \frac{\sum_j \sum_i \delta_{ij}(t)}{\sum_{\substack{n=1 \\ n \neq a}}^{N} \delta_{an}(t)}, \qquad (t_i - t_j) = t, \quad 0 \leq t < T, \quad \forall a,b \in \{1,\ldots,N\}, a \neq b$$

$$\delta_{ij}(t) = \begin{cases} 1, & id(i,t) = id(j,t) \\ 0, & id(i,t) \neq id(j,t) \end{cases}$$



(a)

- Outline
- Motivation
- Proposed system
- Intra-camera tracking
- Time transition distribution
- Spacial distribution estimation
- Advantages
- CIND-CNN
- CUHK-03 dataset
- A more flexible approach
  - Residual learning
  - Triplet network
  - Batch norm.
- Simulations
- Features appearance
- Next step

Advantages

- Achieved context-aware decisions that boost the tracking of people going out-of-view
- More accurate intra-view tracks provided by the strong discrimination capabilities of a deep architecture in re-id
- Re-identifications based on posterior probabilities built from both the spatio-temporal priors over the network
- Automatic and adaptive learning of the logical topology and the time transition relationships of the network
- → Robustness against cameras breakdown

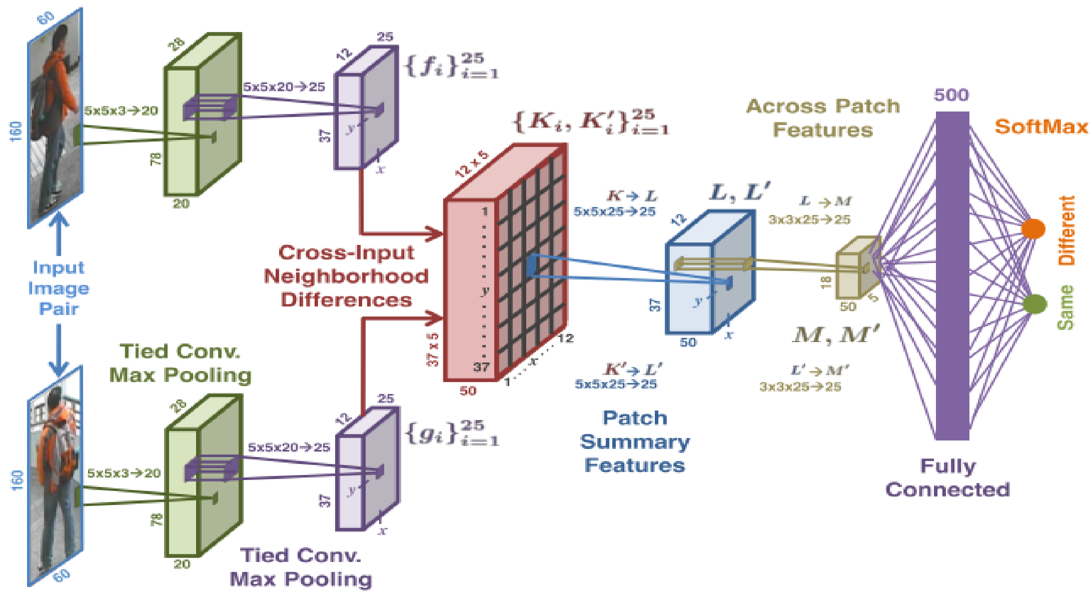
- Outline
- Motivation
- Proposed system
- Intra-camera tracking
- Time transition distribution
- Spacial distribution estimation
- Advantages
- CIND-CNN
- CUHK-03 dataset
- A more flexible approach
  - Residual learning
  - Triplet network
  - Batch norm.
- Simulations
- Features appearance
- Next step

1st CNN implemented

- Outline
- Motivation
- Proposed system
- Intra-camera tracking
- Time transition distribution
- Spacial distribution estimation
- Advantages
- CIND-CNN
- CUHK-03 dataset
- A more flexible approach
  - Residual learning
  - Triplet network
  - Batch norm.
- Simulations
- Features appearance
- Next step

HERIOT WATT UNIVERSITY

1st CNN: Cross-Input Neighborhood Differences CNN



- Each output a_j can be interpreted of the softmax function in terms of predicted probability $p_j=P(y=j|\mathbf{x})$ for the j_{th} class given a sample vector $\mathbf{x}$:

$$P(y=j\,|\,x)=\frac{e^{x^T w_j}}{\sum_{k=1}^{K} e^{x^T w_k}} \qquad L(p,y)=-\log(p_y)$$

- Outline
- Motivation
- Proposed system
- Intra-camera tracking
- Time transition distribution
- Spacial distribution estimation
- Advantages
- CIND-CNN
- CUHK-03 dataset
- A more flexible approach
  - Residual learning
  - Triplet network
  - Batch norm.
- Simulations
- Features appearance
- Next step

Roke
Part of the Chemring Group

Data augmentation and data balancing (minibatches)

- Applying label-preserving operations: random 2D translational transforms on each pedestrian image
- Uncovered stripes of the bounding-box filled with pixels randomly selected from the original image

- First, the gradient of the loss over a mini-batch is an estimate of the gradient over the training set, whose quality improves as the batch size increases.
- Second, computation over a batch can be much more efficient than m computations for individual examples, due to the parallelism afforded by the modern computing platforms.
- Minibatches size: 256 images

- Outline
- Motivation
- Proposed system
- Intra-camera tracking
- Time transition distribution
- Spacial distribution estimation
- Advantages
- CIND-CNN
- CUHK-03 dataset
- A more flexible approach
  - Residual learning
  - Triplet network
  - Batch norm.
- Simulations
- Features appearance
- Next step

CIND-CNN limitations

- Issue: huge peak (~1e20) within the first epoch after some mini-batch iterations

- BP+SGD make it very sensible to initialization values and to the initial learning rate value
- Not very deep
- Deep learning paradigm violation: the function approximated is constrained at the level of the difference layer
- This CNN performs feature extraction and classification by a fully connected layer preventing to make sense of how the features are getting distributed in their space

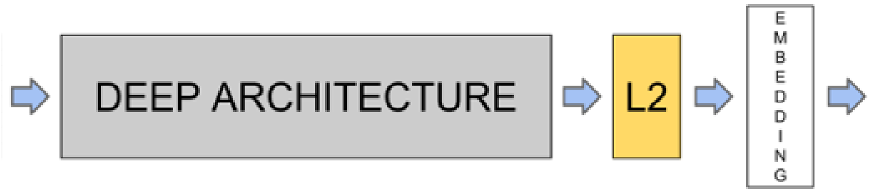
- Outline
- Motivation
- Proposed system
- Intra-camera tracking
- Time transition distribution
- Spacial distribution estimation
- Advantages
- CIND-CNN
- CUHK-03 dataset
- A more flexible approach
  - Residual learning
  - Triplet network
  - Batch norm.
- Simulations
- Features appearance
- Next step

2nd CNN implemented

- Outline
- Motivation
- Proposed system
- Intra-camera tracking
- Time transition distribution
- Spacial distribution estimation
- Advantages
- CIND-CNN
- CUHK-03 dataset
- A more flexible approach
  - Residual learning
  - Triplet network
  - Batch norm.
- Simulations
- Features appearance
- Next step

A more flexible approach

- The end-to-end neural network can learns an optimal metric for discriminating the target automatically.
- This scheme allows to have a clear objective function and to treat the feature maps as multidimensional points in a geometrical (Euclidean) space thus allowing to learn useful representations by distance comparisons



- Advantage: ease of application of any clustering algorithm to associate these "points" exploring the feature space

- Outline
- Motivation
- Proposed system
- Intra-camera tracking
- Time transition distribution
- Spacial distribution estimation
- Advantages
- CIND-CNN
- CUHK-03 dataset
- A more flexible approach
  - Residual learning
  - Triplet network
  - Batch norm.
- Simulations
- Features appearance
- Next step

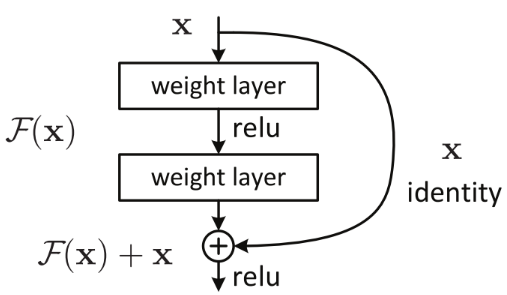
Roke
Part of the Chemring Group

Going deeper by deep residual learning [6]

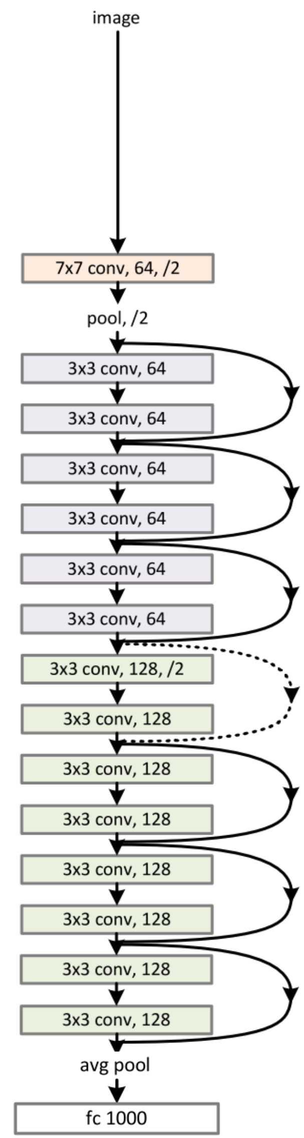
Does a deep CNN learn more the more layers are stuck?

- Problem: vanishing/exploding gradients
  - ✓ This can be addressed by intermediate normalization layers and using Rectified LinearUnits
- Problem: accuracy degradation not caused by overfitting because the training error increases
  - ✓ Deep residual learning framework

- Layers learn residual functions with reference to their inputs instead of learning unreferenced functions.
- Residual networks are easier to optimize.
- They can gain accuracy from increased depth (3.57% error on the ImageNet with 152-layers residual nets)
- Lower complexity at parity of depth: identity shortcuts are parameter-free and this helps the training

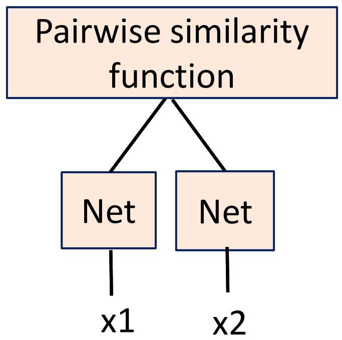


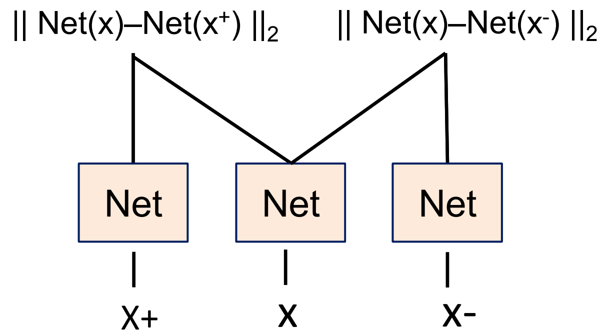
$$\mathcal{F}(\mathbf{x}) := \mathcal{H}(\mathbf{x}) - \mathbf{x}$$



- Outline
- Motivation
- Proposed system
- Intra-camera tracking
- Time transition distribution
- Spacial distribution estimation
- Advantages
- CIND-CNN
- CUHK-03 dataset
- A more flexible approach
  - Residual learning
  - Triplet network
  - Batch norm.
- Simulations
- Features appearance
- Next step

Siamese vs triplet networks





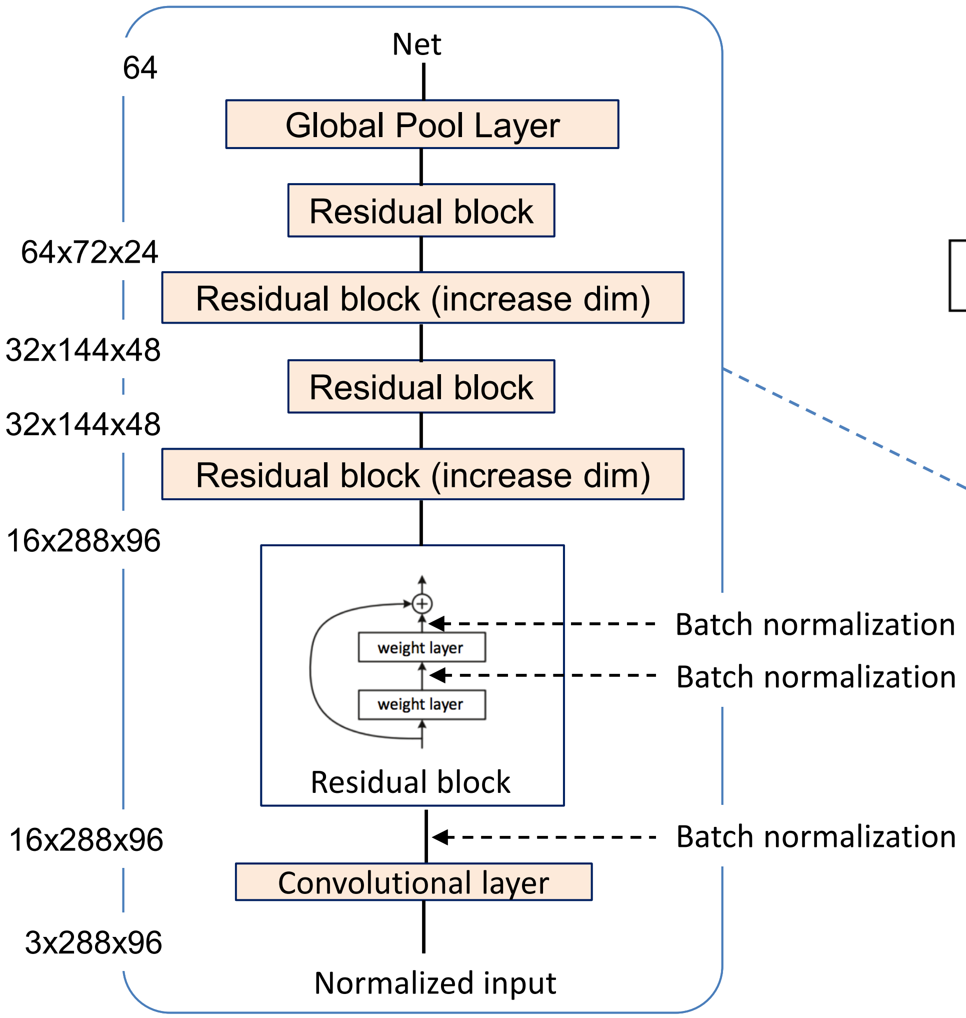
- Siamese networks are sensitive to calibration in the sense that the notion of similarity vs dissimilarity requires context.
- For example, a person might be deemed similar to another person when a dataset of random objects is provided, but might be deemed dissimilar with respect to the same other person when we wish to distinguish between two individuals in a set of individuals only. With the triplet model, such a calibration is not required.
- Triplet networks learns a better representation than siamese networks, improving the classification accuracy in several problems

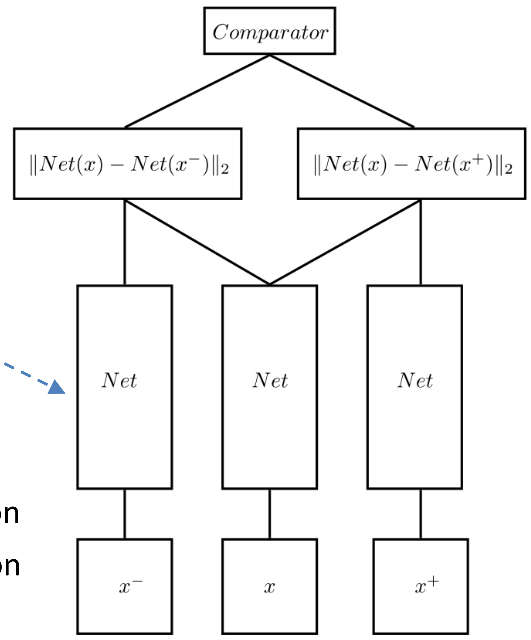
- Outline
- Motivation
- Proposed system
- Intra-camera tracking
- Time transition distribution
- Spacial distribution estimation
- Advantages
- CIND-CNN
- CUHK-03 dataset
- A more flexible approach
  - Residual learning
  - Triplet network
  - Batch norm.
- Simulations
- Features appearance
- Next step

Roke
Part of the Chemring Group

HERIOT WATT UNIVERSITY

2nd CNN: network structure

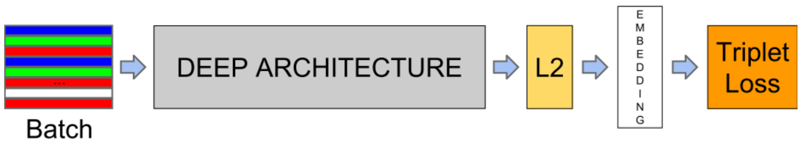




- Outline
- Motivation
- Proposed system
- Intra-camera tracking
- Time transition distribution
- Spacial distribution estimation
- Advantages
- CIND-CNN
- CUHK-03 dataset
- A more flexible approach
  - Residual learning
  - Triplet network
  - Batch norm.
- Simulations
- Features appearance
- Next step

Roke
Part of the Chemring Group

Training by the triplet network scheme



- Learns a mapping into an Euclidean space for identity verification where distances directly correspond to a measure of the similarity of two pedestrians.
- The triplet loss enforces a margin between each pair of images from one person to all other people.

$$\|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2$$

$$\forall \left(f(x_i^a), f(x_i^p), f(x_i^n)\right) \in \mathcal{T}$$

- The loss to minimize is: $$\sum_i^N \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha\right]_+$$
- The Triplet Loss minimizes the distance between an anchor and a positive, both of which have the same identity, and maximizes the distance between the anchor and a negative of a different identity.

- Outline
- Motivation
- Proposed system
- Intra-camera tracking
- Time transition distribution
- Spacial distribution estimation
- Advantages
- CIND-CNN
- CUHK-03 dataset
- A more flexible approach
  - Residual learning
  - Triplet network
  - Batch norm.
- Simulations
- Features appearance
- Next step

Roke
Part of the Chemring Group

Batch normalization (BN)

- Internal Covariate Shift: the change in the distribution of network activations due to the change in network parameters during training.
- The layers need to continuously adapt to the new distribution
- Small changes to the network parameters amplify as the network becomes deeper
- Impact: it slows down the training by requiring lower learning rates and careful parameter initialization

- Normalize each scalar feature independently and add two scale and translation parameters to make it an identity tranform

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma \widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

- It allows to use much higher learning rates and be less careful about initialization
- It acts as a regularizer, often eliminating the need for Dropout
- It achieves the same accuracy with fewer training steps (even for non-decorrelated features)

- Outline
- Motivation
- Proposed system
- Intra-camera tracking
- Time transition distribution
- Spacial distribution estimation
- Advantages
- CIND-CNN
- CUHK-03 dataset
- A more flexible approach
  - Residual learning
  - Triplet network
  - Batch norm.
- Simulations
- Features appearance
- Next step

Roke
Part of the Chemring Group

From simulations…

- Outline
- Motivation
- Proposed system
- Intra-camera tracking
- Time transition distribution
- Spacial distribution estimation
- Advantages
- CIND-CNN
- CUHK-03 dataset
- A more flexible approach
  - Residual learning
  - Triplet network
  - Batch norm.
- Simulations
- Features appearance
- Next step

Roke
Part of the Chemring Group

From simulations…

Augmentation factor 3

- Number of images after augmentation: 42086
- 11 conv layers → ~80000 parameters

Dataset split into three partitions:

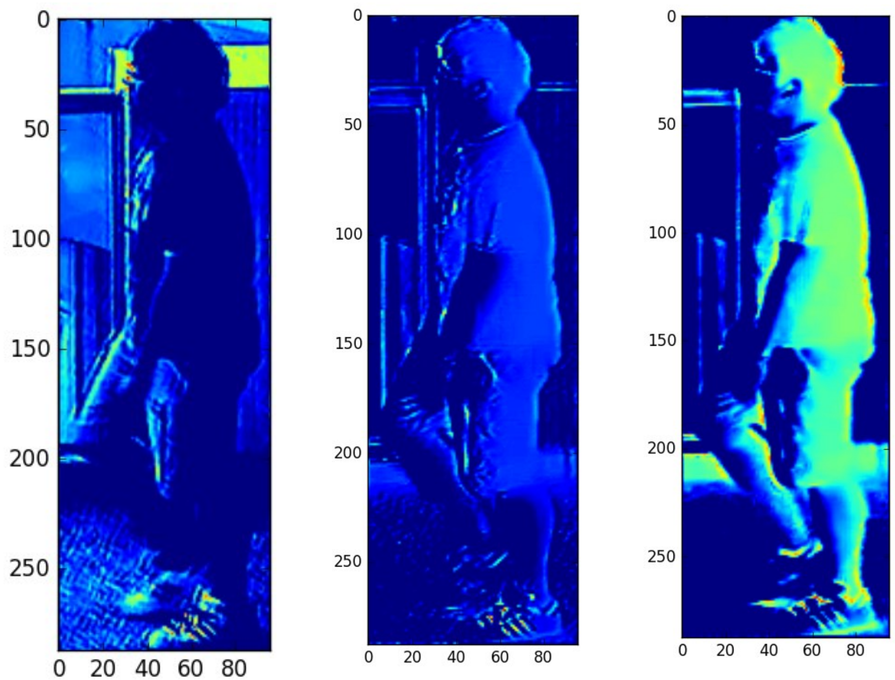
- Training set: 554223 positive (triplet) samples
- Test set: 43500 (triplet) samples (100 identities)
- Validation set: 43500 (triplet) samples (100 identities)

- Depending on the number of parameters of the CNN the training time for each epoch is ~1h 30min
- For each epoch a validation step is also performed for stopping the training when the validation accuracy curve starts decreasing
- Training loss decreasing
- Validation and test accuracy still equal to zero → under investigation

- Outline
- Motivation
- Proposed system
- Intra-camera tracking
- Time transition distribution
- Spacial distribution estimation
- Advantages
- CIND-CNN
- CUHK-03 dataset
- A more flexible approach
  - Residual learning
  - Triplet network
  - Batch norm.
- Simulations
- Features appearance
- Next step

Appearance of Features at each layer

Feature maps extracted at the 1st layer by different filters to be trained:

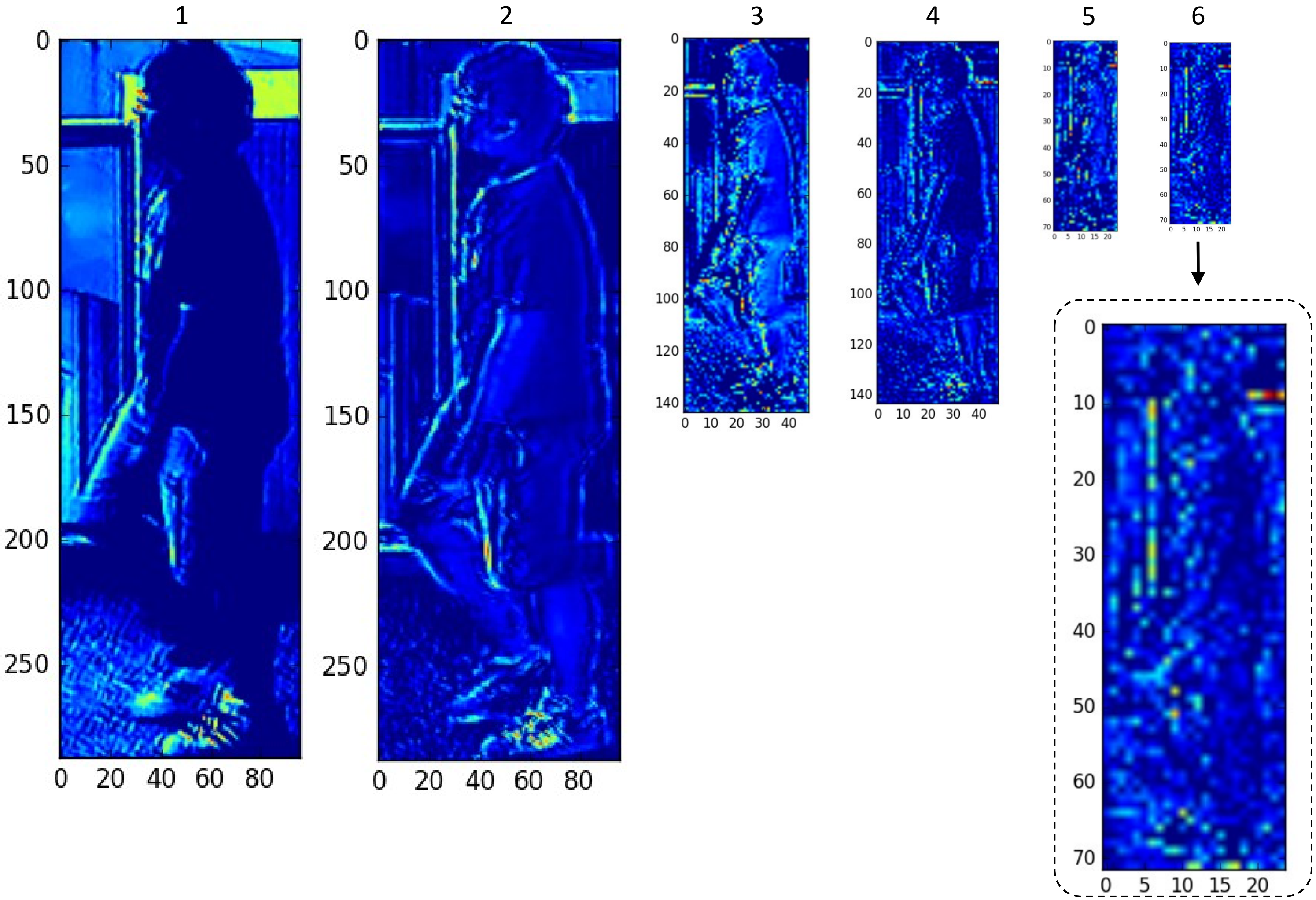


Filter 1 Filter 2 Filter 3

- Outline
- Motivation
- Proposed system
- Intra-camera tracking
- Time transition distribution
- Spacial distribution estimation
- Advantages
- CIND-CNN
- CUHK-03 dataset
- A more flexible approach
  - Residual learning
  - Triplet network
  - Batch norm.
- Simulations
- Features appearance
- Next step

Appearance of Features at each layer

Feature of the same input image extracted at different layers of the CNN in correspondence of the first filter:



- Outline
- Motivation
- Proposed system
- Intra-camera tracking
- Time transition distribution
- Spacial distribution estimation
- Advantages
- CIND-CNN
- CUHK-03 dataset
- A more flexible approach
  - Residual learning
  - Triplet network
  - Batch norm.
- Simulations
- Features appearance
- Next step

Next steps

- Set a suitable number of layers/parameters to achieve state-of-the-art performance in training/testing against CUHK-03 dataset
- Test the performances of the trained CNN gainst SAIVT-BIO video dataset
- Exploring the feature space and apply clustering in the metric space of the representation

- Outline
- Motivation
- Proposed system
- Intra-camera tracking
- Time transition distribution
- Spacial distribution estimation
- Advantages
- CIND-CNN
- CUHK-03 dataset
- A more flexible approach
  - Residual learning
  - Triplet network
  - Batch norm.
- Simulations
- Features appearance
- Next step

Roke
Part of the Chemring Group

References

[1] E. Ahmed, A. V Williams, C. Park, M. Jones, and T. K. Marks, "An Improved Deep Learning Architecture for Person Re-Identification."

[2] Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A Unified Embedding for Face Recognition and Clustering. Retrieved from http://arxiv.org/abs/1503.03832

[5] Yi, D., Lei, Z., Liao, S., & Li, S. Z. (2014). Deep Metric Learning for Person Re-identification. 2014 22nd International Conference on Pattern Recognition, (1), 34–39. http://doi.org/10.1109/ICPR.2014.16

[6] Technologii, C. H., Poc, S., & Multime, G. a. (2013). Deep Residual Learning for Image Recognition, 7(3), 171–180.

[7] Hoffer, E., & Ailon, N. (2014). Deep metric learning using Triplet network, (2010), 1–8. Retrieved from http://arxiv.org/abs/1412.6622

[8] Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv*. Retrieved from http://arxiv.org/abs/1502.03167

[9] Kingma, D., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [Cs], 1–15. Retrieved from http://arxiv.org/abs/1412.6980 http://www.arxiv.org/pdf/1412.6980.pdf

- Outline
- Motivation
- Proposed system
- Intra-camera tracking
- Time transition distribution
- Spacial distribution estimation
- Advantages
- CIND-CNN
- CUHK-03 dataset
- A more flexible approach
  - Residual learning
  - Triplet network
  - Batch norm.
- Simulations
- Features appearance
- Next step

Thank you!

Questions?