Comparing Visual Feature Coding for Learning Disjoint Camera Dependencies

Xiatian Zhu[1]
xiatian.zhu@eecs.qmul.ac.uk
Shaogang Gong[1]
sgg@eecs.qmul.ac.uk
Chen Change Loy[2]
ccloy@visionsemantics.com

[1] School of Electronic Engineering and Computer Science,
Queen Mary, University of London,
London E1 4NS, UK

[2] Vision Semantics,
London E1 4NS, UK

Problem: This work systematically investigates the effectiveness of various visual feature coding schemes for facilitating the learning of time-delayed dependencies among disjoint multi-camera views.

Related work: Quite a few studies [3, 4, 6] have been proposed to model inter-camera dependency across non-overlapping camera views. Learning time-delayed correlations among disjoint cameras in crowded public scenarios is a non-trivial task: (1) *the time gaps between camera views are unknown* therefore activities in two related views may occur at arbitrary time delays with high uncertainty; (2) *the features are inevitably noisy, ambiguous, and may vary drastically* across views owning to illumination condition, camera angles, and changes in object pose. Most state-of-the-art methods typically hand pick a few features tailored to the target environment, with the hope that those chosen features contain robust and sufficient statistics for correlating the time-delayed activity patterns across disjoint views. These manual approaches to hard selection of features are neither principled nor generalisable to different scene context.

Our solution: In this study, we wish to examine the concept that visual features should be coded and selected automatically for robust and accurate time-delayed dependency learning. The contributions of this study are two-fold: (1) We present a systematic study and evaluation to investigate the effectiveness of supervised and unsupervised feature coding methods to facilitate the learning of inter-camera activity pattern dependencies. (2) We systematically evaluate the sensitivity of inter-camera time delayed dependency learning given different training video sizes and region decomposition qualities. These factors are critical for accurate dependency learning but have been largely ignored by the published existing work in the literature.

Approach overview: We employ the Random Forest [2] as the supervised feature coding approach. In particular, given a set of localised features extracted from a region, together with people count training label over time, we first train a regression forest to learn the non-linear mapping between the crowd density and the corresponding low-level features. Given unseen data, we then construct a time series based on the predicted crowd density $\hat{y}$ obtained from the regression forest (*RF pred*), the tree-structured code (*tree code*) [5], or the combination of the two.

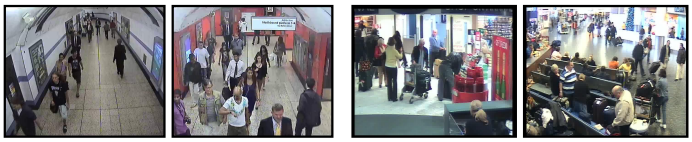
As for unsupervised coding scheme, we use the Latent Dirichlet Allocation (LDA) [1] to map the low-level features into codewords that capture the topic distribution, whereby an image region patch (document) d is treated as a collection of $j = 1 \ldots N_i$ features (words). To form the unsupervised feature codes, given a sequence of localised feature vectors detected from a region, we first perform quantisation on each feature to generate a bag-of-word representation for all image patches. Similar to text documents, these bag-of-word represented image patches are fed into the LDA, which gives us a topic-based representation. Once having the topic-based code (*topic code*), we perform k-means quantisation on them, producing the final compact topic-based code, and concatenate them over time to form a time series.

To solve the problem of using the feature codes for learning inter-camera dependencies, we adopt the Time Delayed Mutual Information (TDMI) proposed in [3] due to its reported effectiveness and simplicity. The input to TDMI are time series generated from either the supervised or the unsupervised coding scheme.

In addition to measuring deviation error in transition time, we propose a new metrics to evaluate the effectiveness of different coding methods, called Mutual Information Margin (MIM):

$$\triangle\mathcal{I} = \frac{\delta(\mathcal{I}_{\text{con}}) - \delta(\mathcal{I}_{\text{uncon}})}{\delta(\mathcal{I}_{\text{con}})}, \delta(\mathcal{I}) = \max(\mathcal{I}) - \min(\mathcal{I}), \quad (1)$$

where $\mathcal{I}_{\text{con}}$ and $\mathcal{I}_{\text{uncon}}$ denote the TDMI function yielded by the connected pairs and unconnected pairs of regions, respectively.



(a) The US dataset (b) The i-LIDS dataset

Figure 1: The example views of the US and the i-LIDS dataset.

Feature Codings	MI-MIM (US)	MI-MIM (i-LIDS)
RF pred	5.1530	7.8577
tree code	-1.7979	-1.7847
RF pred + tree code	-2.3839	-1.0335
topic code	**9.9057**	**16.6349**

Table 1: Sensitivity to the length of the training sequence: the average improvement in MIM of different feature coding methods over the k-means vector quantisation based representation. Mean improved MIM (MI-MIM) was computed by averaging individual percentage of improvement over the testing range.

Feature Codings	MI-MIM (US)	MI-MIM (i-LIDS)
RF pred	10.7670	**13.1541**
tree code	7.8714	2.0040
RF pred + tree code	7.6564	3.5522
topic code	**14.3076**	4.1265

Table 2: Sensitivity to region decomposition: Mean Improved MIM was computed following the same steps as explained in Table 1.

Experiments: We conducted extensive evaluations using two challenging multi-camera datasets: (1) an Underground Station (US) dataset, (2) the i-LIDS Multiple Camera Tracking Scenario (i-LIDS) dataset. See Fig. 1 for example.

The objective of first experiment is to compare the sensitivity of different coding schemes given different lengths of video sequence for time delayed dependency learning (see Table. 1). In the second experiment we evaluated the sensitivity of different coding schemes to the quality of region decomposition. The results are given in Table. 2.

Extensive experiments with both supervised and unsupervised feature coding methods on crowded public scene videos have demonstrated the superiority of the proposed feature coding methods to the conventional k-means vector quantisation, in terms of accuracy in time delayed dependency learning, and robustness to small training sequence size and poor region decomposition quality.

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[3] C. C. Loy, T. Xiang, and S. Gong. Incremental activity modelling in multiple disjoint cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[4] D. Makris, T. Ellis, and J. Black. Bridging the gaps between cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 205–210, 2004.

[5] F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. *Advances in Neural Information Processing Systems*, 19, 2006.

[6] K. Tieu, G. Dalley, and W. E. L. Grimson. Inference of non-overlapping camera network topology by measuring statistical dependence. In *IEEE International Conference on Computer Vision*, pages 1842–1849, 2005.