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

Abstract

This paper systematically investigates the effectiveness of different visual feature coding schemes for facilitating the learning of time-delayed dependencies among disjoint multi-camera views. Accurate inter-camera dependency estimation across non-overlapping camera views is non-trivial especially in crowded scenes where inter-object occlusion can be severe and frequent, and when the degree of crowdedness can change drastically over time. In contrast to existing methods that learn dependencies between disjoint cameras by solely relying on correlating universal object-independent low-level visual features or transition time statistics, we propose to use either supervised or unsupervised feature coding, to establish a robust and reliable representation for estimating more accurately inter-camera activity pattern dependencies. We show comparative experiments to demonstrate the superiority of robust feature coding for learning inter-camera dependencies using benchmark multi-camera datasets of crowded public scenes.

1 Introduction

Disjoint surveillance cameras with non-overlapping field of view (FOV) are typically deployed to monitor a wide-area complex scene. In most cases the statistical dependencies and time gaps among multiple networked cameras are unknown. Discovering these time-delayed dependencies or spatio-temporal correlations is of great benefits to many real-world problems such as topology inference [23, 31], multi-camera tracking [10, 11], person re-identification [19], and global anomaly detection [20, 33].

Learning time-delayed correlations among disjoint cameras 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. Consequently, state-of-the-art methods typically hand pick a few features tailored to the target environment, e.g. foreground features [20], object appearance [11] or transition time statistics [23], 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.

Human cognitive learning to associate objects or relate events is different: apart from learning and abstracting from low-level visual features, they also make use of high-level criterion or description to code low-level features and to resolve ambiguity and uncertainty [2, 27]. This can be seen as *supervised feature coding*. In addition, the human visual system also inherently relies on co-occurrence statistics to establish more reliable representation [8]. This can be treated as *unsupervised feature coding*.

We wish to examine the concept that the features should be coded and selected automatically for robust and accurate time-delayed dependency learning. To this end, we propose to exploit a random forest and a topic model respectively as supervised and unsupervised mechanisms for robust feature coding and implicit feature selection. This is in contrast to existing studies [10, 20, 23, 31] that only utilise predefined object-independent low-level features. The examined coding approaches are flexible in the use of different low-level features, and the methods are scalable to very crowded scenarios. The effectiveness of different coding approaches are demonstrated using two multiple camera datasets, both of which feature complex activity patterns and crowded scene contexts. 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.

2 Related Work

Most existing approaches to event analysis and correlation modelling are devoted to single camera view situations [1, 9, 28, 29]. Extending these methods to scenarios with multiple disjoint cameras is non-trivial due to the unknown inter-camera time gaps and appearance variations across camera views.

There exists a few multi-camera-based approaches that attempt to address the aforementioned problems. These methods can be broadly grouped into two classes: correspondence based [10, 11] and correspondence-free [20, 23, 31] approaches. The method in [10] assumes that visual features of the same objects extracted across views are reliable and robust to associate inter-camera trajectories. In many cases this assumption is not valid due to the significant feature variations across camera views caused by diverse camera angles, the potential changes in illumination and target pose. To address this problem, [23, 31] propose correspondence-free methods by modelling the transition time between disappearance and appearance events observed in different views. However, these models fail to deal with crowded scenes where object detection and tracking become extremely unreliable. This limitation is mitigated to some extent by the work of Loy et al. [20] through estimating the dependency of regional activity time series only using static and moving foreground pixels without object tracking.

All the aforementioned dependency learning algorithms assume that the manually selected features are robust and reliable for their specific datasets. There is no guarantee those selected features can generalise well to other environments. Moreover, they also assume that the same set of features are almost identically effective for all camera views, if not com-

pletely, without considering any specific visual context exhibited in different camera views. For instance, Loy et al. [20] manually select static and moving foreground features for dependency learning for all the views in a complex underground station, ignoring the possible difference of crowd structure in different views. In contrast, the proposed framework in this study does not assume specific features for specific scene. Specifically, we aim to extract a bank of low-level features, and perform feature coding and implicit feature selection in each camera view for obtaining reliable feature representation driven by the context of the target scene.

Different feature coding schemes have been proposed in the past, such as texton [15], visual words [16, 26], discriminative visual codebook [12, 24], or sparse codebook [34]. The introduction of such visual vocabularies has allowed significant advances in image classification and object recognition. Nonetheless, the use of visual coding has not been studied systematically for multi-camera dependency learning. To the best of our knowledge, this is the first study that applies feature coding to generate robust time-series representation for learning inter-camera dependencies.

3 Methodology

We examine the use of feature codes (or visual words) induced from either a supervised random forest learning or an unsupervised topic model clustering for more accurate time delayed dependency learning. In this section, we first present the feature coding approaches (Sec. 3.1), and then describe how our model utilises the feature codes to infer inter-camera activity pattern dependencies (Sec. 3.2).

3.1 Visual Feature Coding

Supervised Feature Coding using Random Forest: Generating visual codebook using a random forest, an ensemble of decision trees [4], has shown promise for visual recognition [14, 24] and pattern recognition problems [32]. In particular, the random forest (RF) is reported to outperform the conventional k-means vector quantisation [7, 30] in terms of training time, memory, testing time, and classification accuracy [24]. In this study, we generate a tree-structured code from low-level features using the random forest, of which the tree construction is driven by top-down localised crowd density in a region for time delayed dependency learning.

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. In particular, we optimise an energy function, which often takes a form of information gain, over a given training set and the associated values of target crowd density. Given a new observation $\mathbf{x}$, a mean prediction is computed by finding the maximum of averaged posterior distributions of all the trees, i.e.

$$\hat{y} = \underset{y}{\operatorname{argmax}} \frac{1}{N_t} \sum_{t}^{N_t} p_t(y|\mathbf{x}), \tag{1}$$

where N_t is the total number of trees in the forest, $p_t(y|\mathbf{x})$ is the posterior of t-th tree.

To generate the supervised feature codes, we follow an approach similar to that proposed in [14, 24, 32]. In particular, given a raw feature vector $\mathbf{x}$, each tree in a random forest produces a binary code of which the length equals to the number of the leaf nodes of that

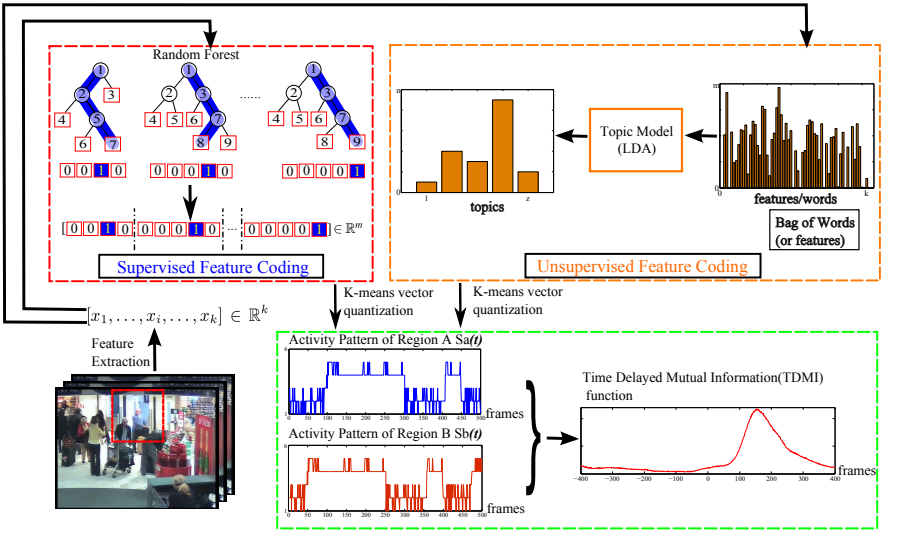


Figure 1: An overview of feature coding comparison for learning inter-camera dependencies.

tree. The binary code has a value one corresponding to the leaf if the feature vector falls into it, and value zero otherwise (see Fig. 1 for example). A random forest with an ensemble of N_t decision trees therefore produces N_t such codes for that feature vector. A tree-structured code is simply a concatenation of binary codes obtained from each tree in the forest. After obtaining the tree-structured code, we perform k-means vector quantisation to generate a more compact final tree-structured code. It is worth pointing out that the code generation benefits from the implicit feature selection mechanism [32] in a random forest. The tree-structured code is thus more robust to non-informative and noisy features.

Unsupervised Feature Coding using Topic Model: In addition to supervised feature coding, we also explore an unsupervised feature coding method using a topic model, which is traditionally used for text mining to discover topics from text documents based on co-occurrence of words.

In this study we employ the Latent Dirichlet Allocation (LDA) [4] to map the low-level features into codewords that capture the topic distribution. In particular, an image region patch (document) d is treated as a collection of $j = 1 \ldots N_i$ features (words). To generate a feature, a topic probability distribution (e.g. multinomial distribution over words) is first chosen from the document multinomial distribution $Multi(\theta_m)$. Then a feature $w_{i,j}$ is sampled from the chosen topic distribution $p(w_{i,j}|\phi_{y_{i,j},n})$. Finally, an image patch can be represented as a combination of topics $z_m, m = 1,2,\ldots,M$, where M is the number of topics in LDA model. Each topic represents a clustering of co-occurring words in all documents.

To form the unsupervised feature codes, we adopt the approach to topic modelling in text analysis. Specifically, 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 (e.g. multinomial probability distribution) for each image patch. The LDA with M topics thus creates a topic-based code, a M-d vector, for the input low-level features in the form of bag-of-word. Once having the topic-based code, like the tree-structured code, we again perform k-means quantisation

on them, producing the final compact topic-based code. Compared to raw low-level features, topic-based code is less sensitive to noise due to its characteristics of representing co-occurring features as clusters [8, 34], having the similar merit of feature selection.

3.2 Time Delayed Dependency Inference

In this section we consider the problem of using the feature codes for learning inter-camera dependencies. The inter-camera dependency can be solved by various algorithms [20, 22, 31]. We adopt the Time Delayed Mutual Information (TDMI) proposed in [20] due to its reported effectiveness and simplicity. Note that the focus of this study is to validate and analyse the effectiveness of feature coding approaches, and not to compare different dependency learning algorithms.

Time Series Construction: The input to the TDMI is represented as a time series $\mathbf{s}_{i,j} = (s_{i,j,1}, \ldots, s_{i,j,t}, \ldots)$, where $s_{i,j,t}$ refers to the feature code within the j-th region of i-th camera view at time t. Different time series can be constructed following different coding schemes described in Sec. 3.1. In particular, for the supervised random-forest based feature coding, we built time series based on the predicted crowd density $\hat{y}$ (*RF pred*), the tree-structured codes (*tree code*), and the combination of the two. As for the unsupervised topic-model based feature coding, we transformed the topic-based codes (*topic code*) into time series.

TDMI Analysis: TDMI [20] was proposed to learn the dependencies among activity patterns observed in a network of cameras. Formally, let two arbitrary regional activity patterns from two camera views be represented as two time series using any type of the aforementioned coding schemes, denoted as $\mathbf{s}_1(t)$ and $\mathbf{s}_2(t)$. The TDMI of $\mathbf{s}_1(t)$ and $\mathbf{s}_2(t+\tau)$ is written as:

$$I(\mathbf{s}_1(t), \mathbf{s}_2(t+\tau)) = \sum_{i=1}^{M_{\mathbf{s}_1}} \sum_{j=1}^{M_{\mathbf{s}_2}} p_{\mathbf{s}_1\mathbf{s}_2}(i,j) \log_2 \frac{p_{\mathbf{s}_1\mathbf{s}_2}(i,j)}{p_{\mathbf{s}_1}(i) p_{\mathbf{s}_2}(j)}, \tag{2}$$

where $M_{\mathbf{s}_1}$ and $M_{\mathbf{s}_2}$ indicate the number of bins of $\mathbf{s}_1(t)$ and $\mathbf{s}_2(t+\tau)$, whilst $p_{\mathbf{s}_1}(\cdot)$ and $p_{\mathbf{s}_2}(\cdot)$ refer to the marginal probability distribution of $\mathbf{s}_1(t)$ and $\mathbf{s}_2(t+\tau)$ separately, and $p_{\mathbf{s}_1\mathbf{s}_2}(\cdot)$ refer to their joint probability distribution. TDMI is a symmetric measurement of dependency between two time series and $I(\mathbf{s}_1(t), \mathbf{s}_2(t+\tau)) \geq 0$, with the equality holding if and only if $\mathbf{s}_1(t)$ and $\mathbf{s}_2(t)$ are independent with each other.

Computing TDMI along different time delays $-T \leq \tau \leq T$ gives us a TDMI function $\mathcal{I}_{\mathbf{s}_1\mathbf{s}_2}(\tau)$ between the two regions:

$$\mathcal{I}_{\mathbf{s}_1\mathbf{s}_2}(\tau) = (I(\mathbf{s}_1(t), \mathbf{s}_2(t-T)), \ldots, I(\mathbf{s}_1(t), \mathbf{s}_2(t+T))). \tag{3}$$

3.3 Evaluation Metrics

We propose the following two metrics to evaluate the effectiveness of coding methods.

Mutual Information Margin (MIM), $\triangle\mathcal{I}$: Most existing methods [19, 20, 23] perform cross-camera correlation or mutual information analysis (see Sec. 3.2) as the initial stage of camera topology inference. For accurate topology inference, the correlation or mutual information function should preserve a high information peak for connected region pair and have a low peak for unconnected region pair. In this study, we introduce a metric to evaluate the quality of the learned correlation or mutual information function. Note that this metric is not

only applicable to TDMI, but can also be used for other inter-camera dependencies function such as cross canonical correlation (xCCA) [19] and cross correlation (xCA) [23]. Mutual Information Margin is defined as:

$$\triangle\mathcal{I} = \frac{\delta(\mathcal{I}_{\text{con}}) - \delta(\mathcal{I}_{\text{uncon}})}{\delta(\mathcal{I}_{\text{con}})}, \delta(\mathcal{I}) = \max(\mathcal{I}) - \min(\mathcal{I}), \tag{4}$$

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. The larger the $\triangle\mathcal{I}$, the more effective the feature representation is in capturing dependencies between connected region pair, whilst suppressing noisy correlation between unconnected region pair. In this study, we compute an averaged MIM using 1 connected pair and 5 random unconnected pairs.

Deviation Error in Transition Time: By transition time we refer to the time gap between an exit event from a camera view to a corresponding entry event in another adjacent camera view by the same individual, an error measurement between the estimation and ground truth used in our experiments is defined as:

$$\varepsilon_{\text{pred}} = \frac{\|T_{\text{pred}} - T_{\text{gt}}\|}{T_{\text{gt}}}, \tag{5}$$

where T_{pred} and T_{gt} represent the most possible normal transition time of prediction and ground truth, separately.

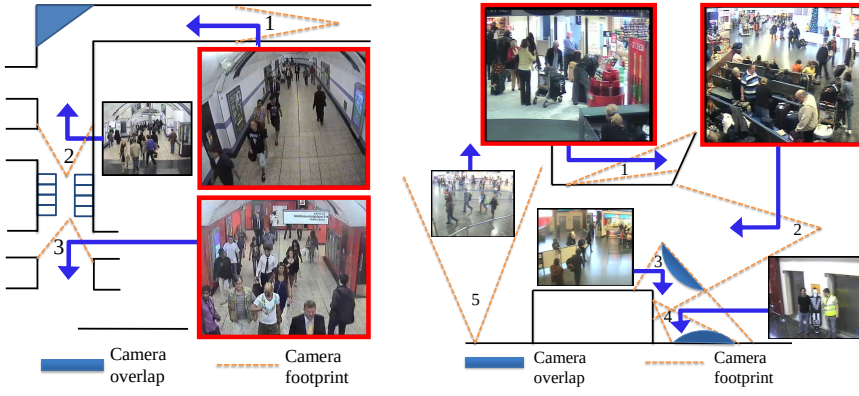
4 Experiments

Datasets: 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. We selected a pair of candidate cameras from each dataset in this study. The layout and example frames of both datasets are given in Fig. 2.

The US dataset (Fig. 2(a)) was recorded from a crowded underground station, with a resolution of 705×577 and fps of 25. This dataset is challenging as (1) there is a large transition gap (average > 1 minute) between two candidate camera views, and (2) there are multiple entrances/exits in the station, which are covered in the views. Both these factors increase the uncertainty in learning the inter-camera time delayed dependency.

The i-LIDS dataset (Fig. 2(b)) was captured with a resolution of 721×577 and 25 fps, from five synchronised and static disjoint cameras installed in a busy airport. The pair of cameras selected for this dataset has a shorter time gap (average < 10 seconds) in comparison to the US dataset. However, unlike the US dataset, the two chosen views of the i-LIDS dataset have very different view fields, i.e. camera 1 has a close view field, whilst camera 2 covers a wider zone with a relatively much farther view. The drastic difference in view fields increases the difficulty in correlating and matching the visual features across views.

Features and Model Settings: We extracted (1) segment-based features including foreground pixels extracted based on static background subtraction [6], moving foreground field by thresholding optical flow magnitude; (2) structural features including edge extracted by a Canny filter; and (3) local texture features based on Local Binary Pattern (LBP) [25]. Note that our framework does not limit the type of features. Different features such as colour histogram or visual attributes [13] can be added without affecting the two coding methods described in Sec. 3.



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

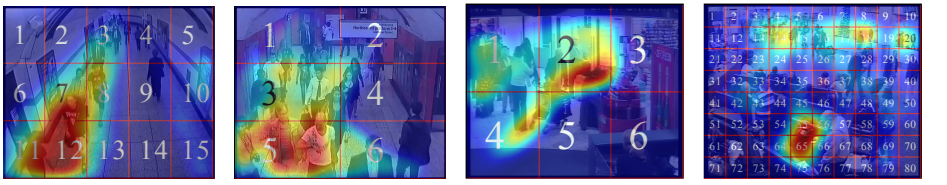
Figure 2: The layout and example views of the US and the i-LIDS dataset. Highlighted views with red boxes are selected for analysis in this study.

The generalisation performance of the random forest is governed by several parameters, one of which is denoted as $\mathcal{T}$, which controls the number of random trees, whilst another parameter N_{minleaf} controls the depth of each tree by limiting the minimum number of training data points falling into each leaf node. In this study, we set $\mathcal{T} = 10$ and $N_{\text{minleaf}} = 50$. As for LDA unsupervised feature coding, we set $M = 5$ roughly corresponding to different degrees of crowdedness, and fixed $N_{\text{it}} = 300$, $\alpha = 10$, and $\beta = 0.01$. These chosen parameter values empirically gave us stable performance.

Region Decomposition and Selection: It is necessary to isolate different activity regions for accurate time delayed dependency learning [19]. Various scene decomposition approaches can be employed [17, 18, 35]. In this study we deliberately choose to segment a scene into equal-sized cells that roughly corresponds to the height of a person observed in the scene. In this way we had the flexibility to alter the region size to evaluate the robustness of different coding schemes to the quality of decomposed regions.

The proposed supervised coding method (Sec. 3) requires people counts over time to learn the crowd structure for code generation. The ideal case would be annotating the head counts at every single region in a scene. However, exhaustive annotation is time-consuming and laborious. In a scene with a single flat ground plane, one can in practice annotate a small region and approximate the crowd count in other regions in the same scene through perspective normalisation. In this study, we selected a region automatically with rich activities based on motion saliency map [21]. We then annotated the head count in that region and extracted perspective normalised features [5]. Despite the approximation, satisfactory people count estimation was obtained. The selected regions for annotation are shown in Fig. 3.

Sensitivity to the Length of Training Sequence: The aim of this experiment is to compare the sensitivity of different coding schemes given different lengths of video sequence for time delayed dependency learning. In general, the longer the sequence the richer the activity information contained in the sequence, thus more accurate time delay and larger MIM can be achieved. However, in most cases such as on-line topology learning [20], the training sample size is limited. Thus it is worthwhile to investigate how fast and accurate a learning algorithm can learn the dependency given limited information. In this experiment we varied the length of the training sequence from 15000 to 30000 frames and measured the time deviation error



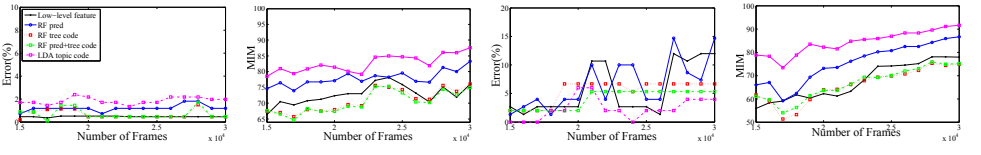
(a) Cam2, US (b) Cam5, US (c) Cam1, i-LIDS (d) Cam2, i-LIDS

Figure 3: Motion Saliency Maps obtained on the US and the i-LIDS dataset. The selected regions are labelled with black digits.

and MIM yielded by different coding methods, i.e. the supervised and unsupervised feature coding approaches (Sec. 3) and the k-means vector quantisation [20] as baseline.

As shown in Fig. 4(a), all coding schemes yielded similar time delay deviation error in the US dataset (< 3% error). Nevertheless, in the i-LIDS dataset, the topic-based codes showed superior and consistent performance given different training sample sizes whilst both the random forest based representation and the k-mean vector quantisation representation yielded larger errors in time delay (see Fig. 4(c)).

As for the MIM (Fig. 4(b) and Fig. 4(d)), it is not surprising to see that both supervised and unsupervised feature coding representations obtained higher MIM than the k-means vector quantisation method. These results suggest that the feature coding is capable of suppressing noisy dependencies between unconnected region pair while capturing inherent activity correlations between connected region pair. In general, the representation based on topic clusterings demonstrated the most favourable performance in both datasets, showing a large performance improvement over the k-means vector quantisation (see Table 1). It is worth pointing out that the person count estimation generated by RF pred (see Eqn.(1)) yielded encouraging results as compared to its more elaborated tree-structured code, suggesting that person count over time can be a useful cue for inter-camera dependency learning.



(a) Time Dev. Err., US (b) MIM, US (c) Time Dev. Err. i-LIDS (d) MIM, i-LIDS

Figure 4: Comparing the sensitivity of different feature coding methods to the video sample size in dependency learning. For the time deviation error, lower is better. For the MIM, higher is better. Figure 4(c-d) use the same legend as Figure 4(a).

Sensitivity to Region Decomposition: In this experiment we evaluated the sensitivity of different coding schemes to the quality of region decomposition. To simulate different sizes of region, we increased the size of a regular cell region from 1/96 of the original frame size to the full frame size for both the US and the i-LIDS dataset. We then inferred the time delayed dependency using different coding schemes and measured the corresponding performance. The results are shown in Fig. 5.

As shown in Fig. 5, in general all coding schemes suffered higher error rate in time delay estimation and lower MIM when the region size was increased. These results are intuitive since by increasing the size of a decomposed region than it should be, one introduced a greater level of noise and redundant information to the region. As a result, it will be much

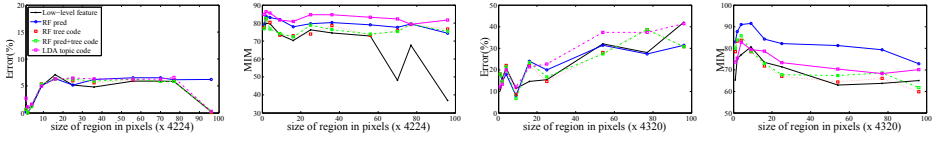
Feature Codings	Mean Improved MIM (US)	Mean Improved MIM (i-LIDS)
RF pred	5.1530	7.8577
tree code	-1.7979	-1.7847
RF pred + tree code	-2.3839	-1.0335
topic code	**9.9057**	**16.6349**

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 was computed by averaging individual percentages of improvement over the testing range.

Feature Codings	Mean Improved MIM (US)	Mean Improved 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.

harder to infer an accurate time delay and mutual information function. Table 2 suggests that both the supervised and unsupervised coding schemes in this study outperformed the conventional k-means quantisation scheme [18] in learning the mutual information function. In particular, the RF pred based codes and topic based codes yielded the best results, suggesting that estimated person density over time and topic clusters are superior in learning inter-camera dependency.



(a) Time Dev. Err., US (b) MIM, US (c) Time Dev. Err., i-LIDS (d) MIM, i-LIDS

Figure 5: Comparing the sensitivity of different feature coding methods to the quality of region decomposition. For the time deviation error, lower is better. For the MIM, higher is better. Figure 5(c-d) use the same legend as Figure 5(a).

5 Conclusion

We have investigated a critical issue that has largely been ignored in existing multi-camera activity analysis studies, i.e. the mechanism of constructing reliable and robust feature representation for learning the time delayed dependencies. In particular, we have presented a supervised coding scheme based on a crowd-sensitive random forest, and an unsupervised coding method based on topic clustering to facilitate more accurate and robust learning of dependency. Extensive experiments 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. Future work includes the investigation of alternative unsupervised and supervised models for feature coding.

References

[1] M. R. Amer and S. Todorovic. A chains model for localizing participants of group activities in videos. In *IEEE International Conference on Computer Vision*, volume 0, pages 786–793, Los Alamitos, CA, USA, 2011.

[2] M. Bar. Visual objects in context. *Nature Reviews: Neuroscience*, 5:617–629, 2004.

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

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

[5] A. Chan, Z.-S. J. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2008.

[6] S. Cohen. Background estimation as a labeling problem. In *IEEE International Conference on Computer Vision*, volume 2, pages 1034–1041, 2005.

[7] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, European Conference on Computer Vision*, volume 1, pages 1–22, 2004.

[8] S. Edelman, H. Yang, B.P. Hiles, and N. Intrator. Probabilistic principles in unsupervised learning of visual structure: human data and a model. *Advances in Neural Information Processing Systems*, 1:19–26, 2002.

[9] U. Gaur, Y. Zhu, B. Song, and A. Roy-Chowdhury. A "string of feature graphs" model for recognition of complex activities in natural videos. In *IEEE International Conference on Computer Vision*, pages 2595–2602, 2011.

[10] O. Javed, Z. Rasheed, K. Shafique, and M. Shah. Tracking across multiple cameras with disjoint views. In *IEEE International Conference on Computer Vision*, pages 952–957, 2003.

[11] O. Javed, K. Shafique, and M. Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 26–33, 2005.

[12] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *IEEE International Conference on Computer Vision*, volume 1, pages 604–610, 2005.

[13] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958, 2009.

[14] V. Lempitsky and A. Zisserman. Learning to count objects in images. In *Advances in Neural Information Processing Systems*, 2010.

[15] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, 43(1):29–44, 2001.

[16] F.-F. Li and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 524–531, 2005.

[17] J. Li, S. Gong, and T. Xiang. Scene segmentation for behaviour correlation. In *European Conference on Computer Vision*, pages 383–395, 2008.

[18] C. C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1988–1995, 2009.

[19] C. C. Loy, T. Xiang, and S. Gong. Time-delayed correlation analysis for multi-camera activity understanding. *International Journal of Computer Vision*, 90(1):106–129, 2010.

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

[21] C. C. Loy, T. Xiang, and S. Gong. Salient motion detection in crowded scenes. In *Special Session on 'Beyond Video Surveillance: Emerging Applications and Open Problems', International Symposium on Communications, Control and Signal Processing*, Invited Paper, 2012.

[22] D. Makris and T. Ellis. Automatic learning of an activity-based semantic scene model. In *IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 183–188, 2003.

[23] 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.

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

[25] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[26] P. Quelhas, F. Monay, J. M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. V. Gool. Modeling scenes with local descriptors and latent aspects. In *IEEE International Conference on Computer Vision*, pages 883–890, 2005.

[27] R. P. Rao, G. Zelinsky, M. Hayhoe, and D. H. Ballard. Eye movements in iconic visual search. *Vision Research*, 42(11):1447–1463, 2002.

[28] M. Rodriguez, J. Sivic, I. Laptev, and J. Y. Audibert. Data-driven crowd analysis in videos. In *IEEE International Conference on Computer Vision*, 2011.

[29] M. S. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *IEEE International Conference on Computer Vision*, 2011.

[30] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision*, pages 1470–1477, 2003.

[31] 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.

[32] C. Vens and F. Costa. Random forest based feature induction. In *IEEE International Conference on Data Mining*, pages 744–753, 2011.

[33] X. Wang, K. Tieu, and W. E. L. Grimson. Correspondence-free activity analysis and scene modeling in multiple camera views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):56–71, 2010.

[34] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1794–1801, 2009.

[35] B. Zhou, X. Wang, and X. Tang. Random field topic model for semantic region analysis in crowded scenes from tracklets. In *IEEE Conference Computer Vision and Pattern Recognition*, 2011.