Video Synopsis by Heterogeneous Multi-Source Correlation

Xiatian Zhu[1], Chen Change Loy[2], Shaogang Gong[1]
[1]Queen Mary, University of London, London E1 4NS, UK
[2]The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
xiatian.zhu@eecs.qmul.ac.uk, ccloy@ie.cuhk.edu.hk, sgg@eecs.qmul.ac.uk

Abstract

Generating coherent synopsis for surveillance video stream remains a formidable challenge due to the ambiguity and uncertainty inherent to visual observations. In contrast to existing video synopsis approaches that rely on visual cues alone, we propose a novel multi-source synopsis framework capable of correlating visual data and independent non-visual auxiliary information to better describe and summarise subtle physical events in complex scenes. Specifically, our unsupervised framework is capable of seamlessly uncovering latent correlations among heterogeneous types of data sources, despite the non-trivial heteroscedasticity and dimensionality discrepancy problems. Additionally, the proposed model is robust to partial or missing non-visual information. We demonstrate the effectiveness of our framework on two crowded public surveillance datasets.

1. Introduction

A critical task in visual surveillance is to automatically make sense of the massive amount of video data by summarising its content using higher-level intrinsic physical events[1] beyond low-level key-frame visual feature statistics and/or object detection counts. In most contemporary techniques, low-level imagery visual cues are typically exploited as the sole information source for video summarisation tasks [11, 17, 6, 12]. On the other hand, in complex and cluttered public scenes there are intrinsically more interesting and relevant higher-level events that can provide a more concise and meaningful summarisation of the video data. However, such events may not be immediately observable visually and cannot be detected reliably by visual cues alone. In particular, surveillance visual data from public spaces is often inaccurate and/or incomplete due to uncontrollable sources of variation, changes in illumination, occlusion, and background clutters [8].

In the context of video surveillance, there are a number of non-visual auxiliary information that can be used to complement the unilateral perspective traditionally offered by visual sources. Examples of non-visual sources include weather report, GPS-based traffic speed data, geo-location data, textual data from social networks, and on-line event schedules. Despite that visual and non-visual data may have very different characteristics and are of different nature, they capture the common physical phenomenon in a scene. This suggests that they are intrinsically correlated, although may be mostly indirect in some latent spaces. Effectively discovering and exploiting such a latent correlation space can bridge the semantic gap between low-level imagery features and high-level semantic interpretation.

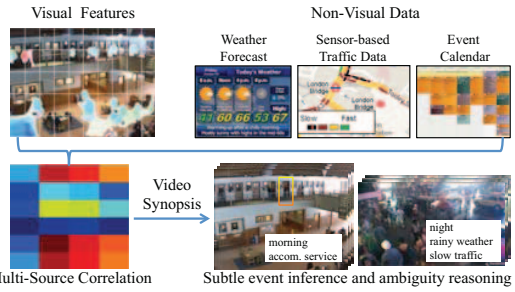

Figure 1. The proposed CC-Forest discovers latent correlations among heterogeneous visual and non-visual data sources, which can be both inaccurate and incomplete, for video synopsis of crowded public scenes.

The objective of this study is to learn a model that associates both visual (*e.g.* optical flow at distributed physical locations) and non-visual (*e.g.* a college event calendar) data for video interpretation and structured synopsis (Fig. 1). The learned model can then be used for event inference and ambiguity reasoning in unseen video data.

Unsupervised mining of latent association and interaction between heterogeneous data sources is non-trivial due to: (1) Disparate sources significantly differ in representation (continuous or categorical), and largely vary in scale and covariance[2]. In addition, the dimension of visual sources often exceeds that of non-visual information to a

[1]Spatiotemporal combinations of human activity and/or interaction patterns, *e.g.* gathering, or environmental state changes, *e.g.* raining or fire.

[2]Also known as the heteroscedasticity problem [4].

1550-5499/13 $31.00 © 2013 IEEE
DOI 10.1109/ICCV.2013.17

great extent (>2000 dimensions of visual features vs. <10 dimensions of non-visual features). Owing to this dimensionality discrepancy problem, a straightforward concatenation of features will result in a representation unfavourably inclined towards the imagery information. (2) Both visual and non-visual data in isolation can be inaccurate and incomplete, especially in surveillance data of public spaces. (3) Non-visual information, *e.g*. event time tables, may not be necessarily available or synchronised with the visual observations. This renders models that expect full and complete input representation impractical. No existing methods are readily applicable to address all the aforementioned challenges in a unified framework.

The main contributions of our study are two-fold. Firstly, we show that coherent and meaningful multi-source based video synopsis can be constructed in an unsupervised manner by learning collectively from heterogeneous visual and non-visual sources. This is made possible by formulating a novel Constrained-Clustering Forest (CC-Forest) with a reformulated information gain function that seamlessly handles multi-heterogeneous data sources dissimilar in representation, distribution, and dimension. Specifically, our model naturally incorporates low-dimensional non-visual data as constraints to high-dimensional visual data. Although both visual and non-visual data in isolation can be inaccurate and incomplete, our model is capable of uncovering and subsequently exploiting the shared latent correlation for video synopsis. As shown in the experiments, combining visual and non-visual data using the proposed method improves the accuracy in video clustering and segmentation, leading towards more meaningful video synopsis. Secondly, the proposed approach is novel in its ability to accommodate partial or completely missing non-visual sources. In particular, in the model training stage, we introduce a joint information gain function that is capable of dynamically adapting to arbitrary number of non-visual constraints. In model deployment, only visual input is required for generating and inferring missing non-visual semantics, relaxing the need for intensive and on-the-fly non-visual information mining.

We demonstrate the effectiveness of our approach on two public surveillance videos. In particular, we demonstrate the usefulness of our framework through generating video synopsis enriched by plausible semantic explanation, providing structured event-based summarisation beyond object detection counts or key-frame feature statistics.

2. Related Work

Contemporary video summarisation methods can be broadly classified into two paradigms, keyframe-based [7, 21, 12] and object-based [18, 17, 6] methods. The keyframe-based approaches select representative keyframes by analysing the low-level imagery properties, *e.g*. object's motion and appearance [7, 12], motion stability of optical flow or global colour differences [21] to generate a storyboard of still images. Object-based techniques [17, 6], on the other hand, rely on object segmentation and tracking to extract object-centered trajectories/tubes, and compress those tubes to reduce spatiotemporal redundancy. Both the above schemes utilise solely visual information and make implicit assumptions about the completeness and accuracy of the visual data available in extracting features or object-centered representations. They are neither suitable nor scalable to complex scenes where visual data are inherently incomplete and inaccurate, mostly the case in typical surveillance videos. Our work differs significantly from these studies in that we exploit not only visual data without object tracking, but also non-visual sources as complementary information in order to discover higher-level events that are visually subtle and difficult to be detected.

Audio information and transcripts have been widely explored for finding highlights in news and broadcast programs [19, 9]. However, these studies are limited to videos recorded in controlled environments. In addition, the complementary sources are well synchronised, mostly noise free and complete as they are extracted from the embedded text metadata. In this work we solve a harder problem. Whilst surveillance videos captured from busy public spaces are typically without auditory signals nor any synchronised transcripts available, we wish to explore alternative non-visual data drawn independently elsewhere from multiple sources, with inherent challenges of being inaccurate and incomplete, unsynchronised to and may also be in conflict with the observed visual data. On a different but related topic, some works have been reported on categorising YouTube videos [22, 20]. In contrast to our problem, these studies enjoy a much wider scope of prior knowledge about correlation of different sources, *e.g*. labelled taxonomy structure, annotated cross-domain sources, together with feedback rating and user-uploaded tags for video clips.

More recently, Huang *et al*. [10] proposed an Affinity Aggregation Spectral Clustering (AASC) method for integrating multiple types of homogeneous information. Their method generates independently multiple affinity matrices via exhaustive pairwise distance computation for every pair of samples of each data source. It suffers from unwieldy representation given high-dimensional data inputs. Importantly, despite it seeks for the optimal weighted combination of the affinity matrices, it does not consider dependency between different data sources in model learning. To overcome these problems, in this work a single affinity matrix that captures correlation between diverse types of sources is derived from a reformulated model of clustering forest. In comparison to [10], our model has a unique advantage in handling missing non-visual data, as shall be demonstrated by extensive experimental evaluations.

3. Video Summarisation from Diverse Sources

We consider the following different sources of information to be taken into account in a multi-source input feature space (Fig. 2-a):

Visual features - We segment a training video into N_v either overlapping or non-overlapping clips, each of which has a duration of T frames. We then extract a d-dimensional visual descriptor from the ith clip denoted by $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,d}) \in \mathbb{R}^d, i = 1, \ldots, N_v$.

Non-visual data - Non-visual data are collected from heterogeneous independent sources. We collectively represent m types of non-visual data associated with the ith clip as $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,m}) \in \mathbb{R}^m, i = 1, \ldots, N_v$. Note that any (or all) dimension(s) of $\mathbf{y}_i$ may be missing.

To facilitate video summarisation with plausible semantic explanation, we need to model latent associations between visual events in video clips and non-visual semantical explanations from independent sources, given a large corpus of video clips and non-visual data. An unsupervised solution is by discovering the natural groupings/clusters from these multiple heterogeneous data sources, so that each cluster represents a meaningful collection of clips with coherent events, associated with unique distributions of non-visual data types. Given a long unseen video, one can then apply a nearest neighbour search in the cluster space to infer the non-visual distribution of any clips in the unseen video.

Discovering coherent heterogeneous data groupings requires the mining of multi-source correlation, which is non-trivial (Sec. 1). A conventional clustering model such as k-means is likely to perform poorly (see experiments in Sec. 5), since the notion of proximity becomes less precise when a single distance function is used for quantifying the groupings of heterogeneous sources differing in representation, distribution, and dimension. In this paper, we address the problem of multi-source correlation and grouping via a *joint optimisation of individual information gains* from different sources, rather than using a 'hard' distance metric for quantification. This naturally isolates the very different characteristics of different sources, thus mitigating the heteroscedastic and dimension discrepancy problems.

A decision forest [3, 23], particularly the clustering forest [1, 13], appears to be a viable solution since its model learning is based on unsupervised information gain optimisation. Nevertheless, the *conventional* clustering forest is not well suited to solving our problem since it expects a full concatenated representation of visual + non-visual sources as input during both the model training and deployment stage. This does not conform to the assumption of only visual data being available during the model deployment for unseen video synopsis. Moreover, in conventional forests, due to the variable selection mechanism, there is no principled way to ensure equal contributions from both visual and non-visual sources in the node splitting process.

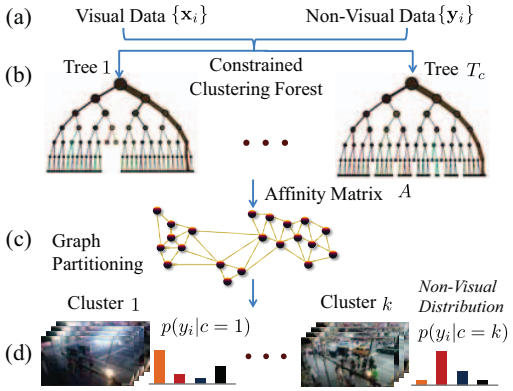

Figure 2. Training steps for learning a multi-source synopsis model.

To overcome the limitations of the conventional clustering forest, we develop a new constrained clustering forest (CC-Forest) by reformulating its optimisation objective function. Figure 2 depicts an overview of the training process of our model.

3.1. Learning Correlation via Information Gain

The proposed CC-Forest, which consists of T_c decision trees (Fig 2-b), is trained similar to a pseudo two-class classification forest [13]. This involves an iterative node splitting procedure that optimises each internal node j via

$$\boldsymbol{\theta}_j^* = \text{argmax}_{\boldsymbol{\theta}_j \in \mathcal{T}} \Delta\mathcal{I}, \tag{1}$$

with a greedy search strategy, where $\boldsymbol{\theta}_j \in \mathcal{T}$ denote the parameters of a test function at the jth split node, and $\Delta\mathcal{I}$ refers to information gain computed as the Gini impurity decrease [2].

In a conventional clustering forest, the information gain $\Delta\mathcal{I}$ is defined as

$$\Delta\mathcal{I} = \mathcal{I}_p - \frac{n_l}{n_p}\mathcal{I}_l - \frac{n_r}{n_p}\mathcal{I}_r, \tag{2}$$

where p, l and r refer to a splitting node, the left and right child node; n denotes the number of samples at a node, with $n_p = n_l + n_r$.

The *main difference* of CC-Forest with respect to the conventional model is that instead of taking a forcefully concatenated visual + non-visual vector as input, it exploits the non-visual information as correlational constraints to guide the tree formation, whilst still using visual features as splitting-variables to grow constrained clustering trees. Specifically, we define a new information gain for node splitting as follow

$$\Delta\mathcal{I} = \underbrace{\alpha_v \frac{\Delta\mathcal{I}_v}{\mathcal{I}_{v0}}}_{\text{visual}} + \underbrace{\sum_{i=1}^{m} \alpha_i \frac{\Delta\mathcal{I}_i}{\mathcal{I}_{i0}}}_{\text{non-visual}} + \underbrace{\alpha_t \frac{\Delta\mathcal{I}_t}{\mathcal{I}_{t0}}}_{\text{temporal}}. \tag{3}$$

Each term in Eqn. (3) is explained as follows:

Visual term - $\Delta\mathcal{I}_v$ denotes the information gain in visual domain. It has a similar derivation as $\Delta\mathcal{I}$ in Eqn. (2), but it is no longer the only factor affecting the node information gain.

Non-visual term - This is a new term we introduce. Specifically, $\Delta\mathcal{I}_i$ denotes the information gain in the ith non-visual data. This new term plays a critical role in that the node splitting is no longer solely dependent on visual data. Instead the mixed information gain in Eqn. (3) encourages data separation not only in the visual domain, but also in the non-visual domains. It is this re-formulation of joint information gain optimisation that provides a chance for associating multiple heterogeneous data sources, and simultaneously balancing the influence exerted by both visual and non-visual information on node splitting.

Temporal term - We also add a temporal smoothness gain $\Delta\mathcal{I}_t$ to encourage temporally adjacent video clips to be grouped together. The information gain of each source is normalised by its initial Gini impurity, denoted by $\mathcal{I}_{v0}$, $\mathcal{I}_{i0}$, and $\mathcal{I}_{t0}$, respectively.

Coping with partial/missing non-visual data - We introduce a new adaptive weighting mechanism to dynamically deal with the inevitable partial/missing non-visual data. Specifically, the coefficients α_v, α_i, and α_t refer to the source weights, with $\alpha_v + \sum_{i=1}^{m} \alpha_i + \alpha_t = 1$. When there are no missing non-visual data, we assume the visual, non-visual, and temporal terms carry equally useful information, we thus set $\alpha_v = 0.5$, and $\alpha_t = \alpha_i = \frac{1-\alpha_v}{m+1}$, with m the number of non-visual sources. In the case of partial/missing non-visual information, suppose the missing proportion of the ith non-visual type in a tree is δ_i, we reduce its weight from α_i to $\alpha_i - \delta_i\alpha_i$. The total reduced weight $\sum_i \delta_i\alpha_i$ will then be distributed evenly to the weights corresponding to all individual sources to ensure $\alpha_v + \sum_{i=1}^{m} \alpha_i + \alpha_t = 1$. This linear adaptive weighting method produces satisfactory performance in our experiments.

3.2. Multi-Source Latent Cluster Discovery

The multi-source feature space is high-dimensional ($>$ 2000 dimensions). This makes learning data structure by clustering computationally difficult. To this end, we consider spectral clustering on manifold to discover latent clusters in a lower dimensional space (Fig 2-c).

Spectral clustering [24] groups data using eigenvectors of an affinity matrix derived from the data. The learned CC-Forest offers an effective way to derive the required affinity matrix. Specifically, each individual tree within the CC-Forest partitions the training samples at its leaves $l(\mathbf{x})$: $\mathbb{R}^d \rightarrow \mathbb{L} \subset \mathbb{N}$, where l represents a leaf index and $\mathbb{L}$ refers to the set of all leaves in a given tree. For each clustering tree, we first compute a tree-level $N_v \times N_v$ affinity matrix A^t with elements defined as $A^t_{i,j} = \exp^{-\text{dist}^t(\mathbf{x}_i,\mathbf{x}_j)}$ with

$$\text{dist}^t(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 0 & \text{if } l(\mathbf{x}_i) = l(\mathbf{x}_j), \\ +\infty & \text{otherwise.} \end{cases} \quad (4)$$

We assign the maximum affinity (affinity=1, distance=0) to points $\mathbf{x}_i$ and $\mathbf{x}_j$ if they fall into the same leaf node, and the minimum affinity (affinity=0, distance=$+\infty$) otherwise. By averaging all tree-level affinity matrices we obtain a smooth matrix as $A = \frac{1}{T_c}\sum_{t=1}^{T_c} A^t$, with $A_{i,i} = 0$.

Subsequently, we symmetrically normalise A to obtain $S = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ where D denotes a diagonal matrix with elements $D_{i,i} = \sum_j A_{i,j}$. Given S, we perform spectral-clustering to discover the latent clusters of training clips with the number of clusters automatically determined [24]. Each training clip $\mathbf{x}_i$ is then assigned to a cluster $c_i \in \mathcal{C}$, with $\mathcal{C}$ the set of all clusters. The learned clusters group similar clips both visually and semantically, each associated with a unique distribution of each non-visual data (Fig. 2-d). We denote the distribution of the ith non-visual data type of the cluster c as $p(y_i|c) \propto \sum_{\mathbf{x}_j \in \mathbf{X}_c} p(y_i|\mathbf{x}_j)$, where $\mathbf{X}_c$ represents the set of training samples in c.

3.3. Structure-Driven Non-Visual Tag Inference

To summarise a long unseen video with high-level interpretation, we need to first infer semantic contents of each clip $\mathbf{x}^*$ in the video. To complete such a task we can exploit the non-visual distributions associated with each cluster discovered (Sec. 3.2). A straightforward way to compute the tag distribution $p(y_i|\mathbf{x}^*)$ of $\mathbf{x}^*$ is to search for its nearest cluster $c^* \in \mathcal{C}$, and let $p(y_i|\mathbf{x}^*) = p(y_i|c^*)$. However, we found this hard cluster assignment strategy susceptible to outliers in $\mathcal{C}$. To mitigate this problem, we propose a more robust approach utilising the CC-Forest tree structures for soft cluster assignment (Fig. 3).

First, we trace the leaf $l_t(\mathbf{x}^*)$ of each tree that $\mathbf{x}^*$ falls into (Fig. 3-a). Second, we retrieve the training samples $\{\mathbf{x}_i\}$ associated with $l_t(\mathbf{x}^*)$ and their cluster membership $C_t = \{c_i\} \subset \mathcal{C}$. Third, within each leaf $l_t(\mathbf{x}^*)$ we search for the nearest cluster c_t^* of $\mathbf{x}^*$ against the centroids of C_t rather than $\mathcal{C}$ (Fig. 3-b), with:

$$c_t^* = \text{argmin}_{c \in C_t} ||\mathbf{x}^* - \mu_c||, \quad (5)$$

with μ_c the centroid of the cluster c, estimated as $\mu_c = \frac{1}{|\mathbf{X}_c|}\sum_{\mathbf{x}_i \in \mathbf{X}_c} \mathbf{x}_i$, where $\mathbf{X}_c$ represents the set of training samples in c.

Once c_t^* is found, we retrieve the associated tag distribution $p(y_i|c_t^*)$. To achieve a smooth prediction, we average all $p(y_i|c = c_t^*)$ obtained from individual trees as (Fig. 3-c):

$$p(y_i|\mathbf{x}^*) = \frac{1}{T_c}\sum_{t=1}^{T_c} p(y_i|c_t^*). \quad (6)$$

A tag of the ith non-visual data type is computed as

$$\hat{y}_i = \text{argmax}_{y_i} \left[p(y_i|\mathbf{x}^*)\right]. \quad (7)$$

With the above steps, we can estimate $\hat{y}_i$ for $i = 1, \ldots, m$. In Sec.5.2, we shall show examples on using the proposed tag inference method (Fig. 3) for generating video synopsis enriched by non-visual semantic labels.

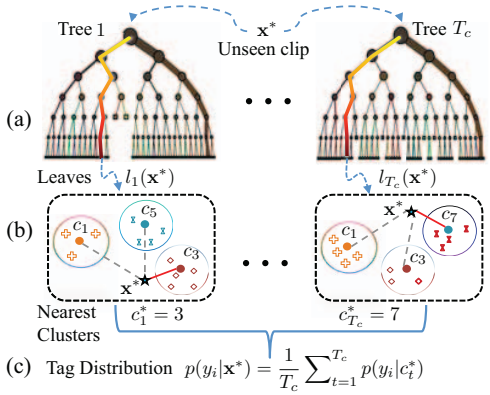

Figure 3. Structure-driven non-visual tag inference: (a) Channel an unseen clip $\mathbf{x}^*$ into individual trees; (b) Estimate the nearest clusters of $\mathbf{x}^*$ within the leaves it falls into: hollow circles denote clusters; (c) Compute the tag distributions by averaging tree-level predictions.

4. Experimental Settings

Datasets - We conducted experiments on two datasets collected from publicly accessible webcams that feature an outdoor and an indoor scene respectively: (1) the TImes Square Intersection (TISI) dataset[3], and (2) the Educational Resource Centre (ERCe) dataset[4]. There are a total of 7324 video clips spanning over 14 days in the TISI dataset, whilst a total of 13817 clips were collected across a period of two months in the ERCe dataset. Each clip has a duration of 20 seconds. The details of the datasets and training/deployment partitions are given in Table 1. Example frames are shown in Fig. 4.

The TISI dataset is challenging due to severe inter-object occlusion, complex behaviour patterns, and large illumination variations caused by both natural and artificial light sources at different day time. The ERCe dataset is non-trivial due to a wide range of physical events involved that are characterised by large changes in environmental setup, participants, crowdedness, and intricate activity patterns.

	Resolution	FPS	# Training Clip	# Test Clip
TISI	550 × 960	10	5819	1505
ERCe	480 × 640	5	9387	4430

Table 1. Details of datasets.

Visual and non-visual sources - We extracted a variety of visual features from each video clip: (a) colour features including RGB and HSV; (b) local texture features based on Local Binary Pattern (LBP) [15]; (c) optical flow; (d) holistic features of the scene based on GIST [16]; and (e) person and vehicle (only on the TISI dataset) detections [5].

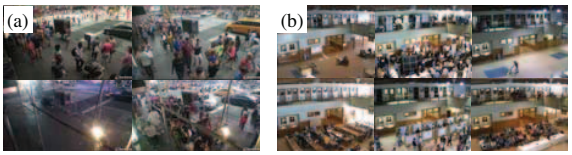

Figure 4. Example views of the (a) TISI and (b) ERCe datasets.

We collected 10 types of non-visual sources for the TISI dataset: (a) weather data extracted from the WorldWeatherOnline[5] including 9 elements: temperature, weather type, wind speed, wind direction, precipitation, humidity, visibility, pressure, and cloud cover; (b) traffic data from the Google Maps with 4 levels of traffic speed: very slow, slow, moderate, and fast. For the ERCe dataset, we collected data from multiple independent on-line sources about the time table of events including: No Scheduled Event (No Schd. Event), Cleaning, Career Fair, Forum on Gun Control and Gun Violence (Gun Forum), Group Studying, Scholarship Competition (Schlr. Comp.), Accommodative Service (Accom. Service), Student Orientation (Stud. Orient.).

Note that other visual features and non-visual data types can be considered without altering the training and inference methods of our model as the CC-Forest can cope with different families of visual features as well as distinct types of non-visual sources.

Baselines - We compare the proposed model Visual + Non-Visual + CC-Forest (*VNV-CC-Forest*) with: (1) *VO-Forest* - a conventional forest [1] trained with visual features alone, to demonstrate the benefits from using non-visual sources [6]. (2) *VNV-Kmeans* - k-means using both visual and non-visual sources, to highlight the heteroscedastic and dimensionality discrepancy problem caused by heterogeneous visual and non-visual data. (3) *VNV-AASC* - a state-of-the-art multi-modal spectral clustering method [10] learned with both visual and non-visual data, to demonstrate the superiority of *VNV-CC-Forest* in handling diverse data representations and correlating multiple sources through joint information gain optimisation. (5) *VPNV(R)-CC-Forest* - a variation of our model but with *R%* of training samples having arbitrary number of partial non-visual types, to evaluate the robustness of our model in coping with partial/missing non-visual data.

Implementation details - The clustering forest size T_c was set to 1000. The depth of each tree is automatically determined by setting the size of the leaf node ϕ, which we fixed to 2 throughout our experiments. We used a linear data separation [3] as the test function for node splitting. We set the

[3] http://www.eecs.qmul.ac.uk/~xz303/downloads_qmul_TISI_dataset.html

[4] http://www.eecs.qmul.ac.uk/~xz303/downloads_qmul_ERCe_dataset.html

[5] http://www.worldweatheronline.com/

[6] Since non-visual data is not available for test clips, so evaluating a forest that takes only non-visual inputs is not possible.

same number of clusters across all methods for a fair comparison. This cluster number was discovered automatically using the method presented in [24] (see Sec 3.2).

5. Evaluations

5.1. Multi-Source Latent Cluster Discovery

For validating the effectiveness of different clustering models for multi-source clustering in order to provide more coherent video content grouping (Sec. 3.2), and to improve the accuracy in non-visual tag inference (Sec. 3.3), we compared the quality of clusters discovered by different methods. We quantitively measured the mean entropy [25] (lower is better) of non-visual distributions $p(\mathbf{y}|c)$ associated with clusters to evaluate how coherent video contents are grouped, with an assumption that all methods have access to all non-visual data during the entropy computation.

Dataset	TISI		ERCe
$p(\mathbf{y}\|c)$	Traffic Speed	Weather	Events
VO-Forest	0.8675	1.0676	0.0616
VNV-Kmeans	0.9197	1.4994	1.2519
VNV-AASC	0.7217	0.7039	0.0691
VNV-CC-Forest	0.7262	**0.6071**	**0.0024**
VPNV10-CC-Forest	**0.7190**	0.6261	**0.0024**
VPNV20-CC-Forest	0.7283	0.6497	0.0090

Table 2. Quantitative comparison on cluster purity using mean entropy.

It is evident from Table 2 that the proposed *VNV-CC-Forest* model achieves the best cluster purity on both datasets. Despite that there are gradual degradations in clustering quality when we increased the non-visual data missing proportion, overall the *VNV-CC-Forest* model copes well with partial/missing non-visual data. Inferior performance of *VO-Forest* to *VNV-CC-Forest* suggests the importance of learning from auxiliary non-visual sources. Nevertheless, not all methods perform equally well when learning from the same visual and non-visual sources: the k-means and AASC perform much poorer in comparison to CC-Forest. The results suggest the proposed joint information gain criterion (Eqn. (3)) is more effective in handling heterogeneous data than conventional clustering models.

For qualitative comparison, we show some examples using the TISI dataset for detecting 'sunny' weather (Fig. 5). It is evident that only the *VNV-CC-Forest* is able to provide coherent video grouping, with only slight decrease in clustering purity given partial/missing non-visual data. Other methods including *VNV-AASC* result in a large cluster either leaving out some relevant ones or including many non-relevant clips, with most of them were under the influence of strong artificial lighting sources. These non-relevant clips are visually 'close' to sunny weather, but semantically not. The *VNV-CC-Forest* model avoids this mistake by correlating both visual and non-visual sources in an information theoretic sense.

Figure 5. Qualitative comparison on cluster quality between different methods on the TISI dataset. A key frame of each video clip is shown. (X/Y) in the brackets - X refers to the number of clips with sunny weather as shown in the images in the first two columns. Y is the total number of clips in a cluster. The frames inside the red boxes refer to those inconsistent clips in a cluster.

(%)	*VO-Forest*	*VNV-Kmeans*	*VNV-AASC*	*VNV-CC-Forest*	*VPNV10-CC-Forest*	*VPNV20-CC-Forest*
Traffic	27.62	37.80	36.13	35.77	37.99	**38.05**
Weather	50.65	43.14	44.37	**61.05**	55.99	54.97

Table 3. Comparison of tagging accuracy on the TISI Dataset.

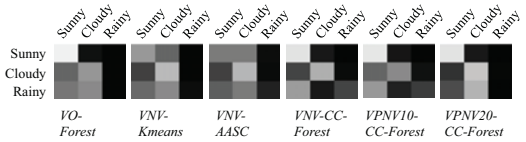

Figure 6. Weather tagging confusion matrices on the TISI Dataset.

5.2. Contextually-Rich Multi-Source Synopsis

Generating video synopsis with semantically meaningful contextual labels requires accurate tag prediction (Sec. 3.3). In this experiment we compared the performance of different methods in inferring tag labels given unseen video clips extracted from long video streams. For quantitative evaluation, we manually annotated three different weathers (sunny, cloudy and rainy) and four traffic speeds on all the TISI test clips, as well as eight event categories on all the ERCe test clips. Note that in the deployment phase the input to all models consists of only visual data.

Correlating and tagging video by weather and traffic conditions - Video synopsis by tagging weather and traffic conditions was tested using the TISI outdoor dataset. It is observed that performance of different methods (Table 3) is largely in line with their performance in multi-source clustering (Sec. 5.1). Further comparisons of their confusion matrices on weather conditions tagging are provided in Fig. 6. It is worth pointing out that *VNV-CC-Forest* not only outperforms other baselines in isolating the sunny weather, but also performs well in distinguishing the visually ambiguous cloudy and rainy weathers. In contrast, both *VNV-Kmeans* and *VNV-AASC* mistake most of 'rainy' scenes as either 'sunny' or 'cloudy', as they can be visually similar.

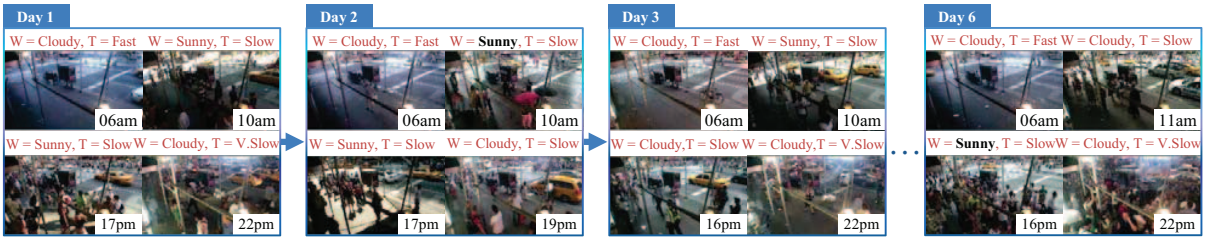

Figure 8. A synopsis with a multi-scale overview of weather+traffic changes over multiple days. Black **bold** prints = failure predictions.

(%)	*VO-Forest*	*VNV-Kmeans*	*VNV-AASC*	*VNV-CC-Forest*	*VPNV10-CC-Forest*	*VPNV20-CC-Forest*
No Schd. Event	79.48	**87.91**	48.51	55.98	47.96	55.57
Cleaning	39.50	19.33	45.80	41.28	**46.64**	46.22
Career Fair	94.41	59.38	79.77	**100.0**	**100.0**	**100.0**
Gun Forum	74.82	44.30	84.93	83.82	**85.29**	**85.29**
Group Studying	92.97	46.25	96.88	**97.66**	**97.66**	95.78
Schlr Comp.	82.74	16.71	89.40	99.46	**99.73**	99.59
Accom. Service	0.00	0.00	21.15	**37.26**	**37.26**	37.02
Stud. Orient.	60.94	9.77	38.87	88.09	**92.38**	88.09
Average	65.61	35.45	63.16	75.69	75.87	**75.95**

Table 4. Comparison of tagging accuracy on the ERCe dataset.

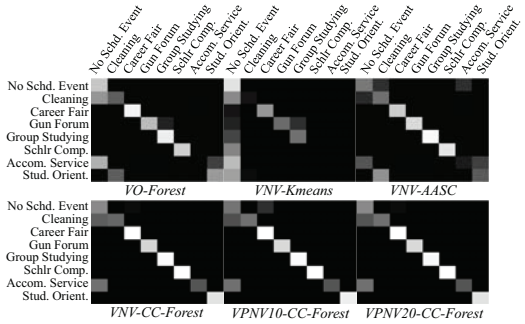

Figure 7. Tag inference confusion matrices on the ERCe dataset.

Figure 9. Summarisation of some key events taking place during the first two months of a new semester on a university campus. The top-left corner numbers in each window are month-date whilst the bottom-right numbers are the hours on a day.

Correlating and tagging video by semantic events - Video synopsis by correlating and tagging higher-level semantic events was tested using the ERCe dataset. The results and the associated confusion matrices are given in Table 4 and Fig. 7 respectively. By *VO-Forest*, poor results are observed especially on 'Accom. Service' event, which involves only subtle activity patterns, *i.e.* students visiting particular rooms located at the first floor. It is evident that using visual information alone is not sufficient to discover such type of event without the support of additional non-visual sources (the semantic gap problem).

Due to the typically high dimension of visual sources in comparison to non-visual sources, the latter is often overwhelmed by the former in representation. *VNV-Kmeans* severely suffers from this problem as most event predictions are biased to the 'No Schd.' event that is more common and frequent visually. This suggests that the conventional distance-based clustering is poor in coping with the inherent heteroscedasticity and dimension discrepancy problems in modelling heterogeneous multi-source independent data. *VNV-AASC* attempts to circumvent this problem by seeking for an optimal weighted combination of affinity matrices derived independently from different data sources. However this is proved challenging, particularly when each source is inherently noisy and ambiguous, leading to an inaccurate combined affinity. In contrast, the proposed *VNV-CC-Forest* correlates different sources via a joint information gain criterion to effectively alleviate the heteroscedasticity and dimension discrepancy problem, leading to more robust and accurate tagging performance. Again, it is observed that *VPNV(10/20)-CC-Forest* performed comparably to *VNV-CC-Forest*, further validating the robustness of *VNV-CC-Forest* in tackling partial/missing non-visual data with the proposed adaptive weighting mechanism (Sec. 3.1). Occasionally *VPNV(10/20)-CC-Forest* even slightly outperforms *VNV-CC-Forest*. We observed that this can be caused by noisy non-visual information. Therefore the missing of some noisy information leads to better results in a few cases.

After inferring the non-visual semantics for the unseen clips, one can readily generate various types of concise video synopsis with enriched contextual interpretation or relevant high-level physical events, using a similar strategy as [14]. We show two examples here. In Fig. 8 we show a synopsis with a multi-scale overview of weather changes and traffic condition over multiple days. Some failure tagging cases are indicated in bold print. In Fig. 9 we depict a synopsis highlighting some of the key events taking place during the first two months of a new semester in a university campus.

5.3. Further Analysis

The superior performance of *VNV-CC-Forest* can be better explained by examining more closely the capability of CC-Forest in uncovering and exploiting the intrinsic association among different visual sources and more critically among visual and non-visual auxiliary sources. This indirect correlation among multi-heterogeneous data sources results in well-structured decision trees, subsequently leading to more consistent clusters and more accurate semantics inference. We show an example here. It is intuitive that vehicle and person counts should correlate in a busy scene like TISI. Our CC-Forest discovered this correlation (see Fig. 10-a), so the less reliable vehicle detection from distance against a cluttered background, could enjoy a latent support from the more reliable person detection in regions 5-16 close to the camera view. Moreover, visual sources also benefited from the correlational support from non-visual information through the cross-source optimisation of individual information gains (Eqn. (3)). For example, in Fig. 10-b, it is evident that the unreliable vehicle detection at far view-field (region 1) is well supported by the traffic-speed non-visual data.

(a) Visual-visual. (b) Vehicle detection and traffic speed.

Figure 10. The latent correlations among heterogeneous visual and multiple non-visual sources discovered on the TISI dataset.

6. Conclusion

We have presented a novel unsupervised method for generating contextually-rich and semantically-meaningful video synopsis by correlating visual features and independent sources of non-visual information. The proposed model, which is learned based on a joint information gain criterion for learning latent correlations among different independent data sources, naturally copes with diverse types of data with different representation, distribution, and dimension. Crucially, it is robust to partial and missing non-visual data. Experimental results have demonstrated that combining both visual and non-visual sources facilitates more accurate video event clustering with richer semantical interpretation and video tagging than using visual information alone. The usefulness of the proposed model is not limited to video summarisation, and can be explored for other tasks such as multi-source video retrieval and indexing. In addition, the semantic tag distributions inferred by the model can be exploited as the prior for other surveillance tasks such as social role and/or identity inference. Future work include how to generalise a learned model to new scenes that are different from the training environments.

References

[1] L. Breiman. Random forests. *ML*, 45(1):5–32, 2001. 3, 5

[2] L. Breiman, J. Friedman, C. Stone, and R. Olshen. *Classification and regression trees*. Chapman & Hall/CRC, 1984. 3

[3] A. Criminisi and J. Shotton. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision*, 7(2-3):81–227, 2012. 3, 5

[4] R. Duin and M. Loog. Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion. *TPAMI*, 26(6):732 –739, 2004. 1

[5] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2010. 5

[6] S. Feng, Z. Lei, D. Yi, and S. Z. Li. Online content-aware video condensation. In *CVPR*, pages 2082–2087, 2012. 1, 2

[7] D. Goldman, B. Curless, D. Salesin, and S. Seitz. Schematic storyboards for video editing and visualization. In *SIGGRAPH*, volume 25, pages 862–871, 2006. 2

[8] S. Gong, C. C. Loy, and T. Xiang. Security and surveillance. In *Visual Analysis of Humans*, pages 455–472. Springer, 2011. 1

[9] Y. Gong. Summarizing audiovisual contents of a video program. *EURASIP J. Appl. Signal Process.*, 2003:160–169, 2003. 2

[10] H.-C. Huang, Y.-Y. Chuang, and C.-S. Chen. Affinity aggregation for spectral clustering. In *CVPR*, pages 773–780, 2012. 2, 5

[11] H. Kang, X. Chen, Y. Matsushita, and X. Tang. Space-time video montage. In *CVPR*, pages 1331–1338, 2006. 1

[12] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, pages 1346–1353, 2012. 1, 2

[13] B. Liu, Y. Xia, and P. S. Yu. Clustering through decision tree construction. In *CIKM*, pages 20–29, 2000. 3

[14] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li. A user attention model for video summarization. In *ACM MM*, pages 533–542, 2002. 7

[15] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *TPAMI*, 24(7):971–987, 2002. 5

[16] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42:145–175, 2001. 5

[17] Y. Pritch, A. Rav-Acha, and S. Peleg. Nonchronological video synopsis and indexing. *TPAMI*, 30(11):1971–1984, 2008. 1, 2

[18] A. Rav-Acha, Y. Pritch, and S. Peleg. Making a long video short: Dynamic video synopsis. In *CVPR*, pages 435–441, 2006. 2

[19] C. Taskiran, Z. Pizlo, A. Amir, D. Ponceleon, and E. Delp. Automated video program summarization using speech transcripts. *TMM*, 8(4):775 –791, 2006. 2

[20] G. Toderici, H. Aradhye, M. Pasca, L. Sbaiz, and J. Yagnik. Finding meaning on YouTube: Tag recommendation and category discovery. In *CVPR*, pages 3447–3454, 2010. 2

[21] B. T. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. *ACM TOMCCAP*, 3(1):3, 2007. 2

[22] Z. Wang, M. Zhao, Y. Song, S. Kumar, and B. Li. YouTubeCat: Learning to categorize wild web videos. In *CVPR*, pages 879–886, 2010. 2

[23] H. Yang and I. Patras. Sieving regression forest votes for facial feature detection in the wild. In *ICCV*, 2013. 3

[24] L. Zelnik-manor and P. Perona. Self-tuning spectral clustering. In *NIPS*, pages 1601–1608, 2004. 4, 6

[25] Y. Zhao and G. Karypis. Empirical and theoretical comparisons of selected criterion functions for document clustering. *ML*, 55(3):311–331, 2004. 6