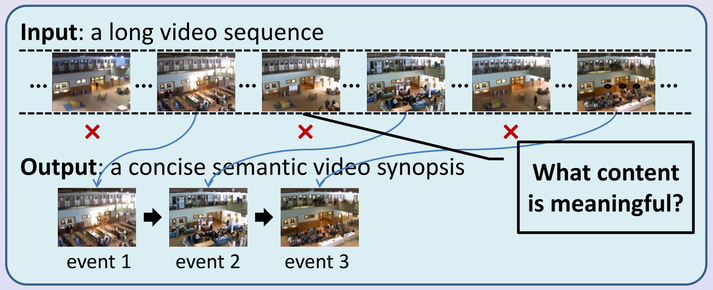
Video Synopsis by Heterogeneous Multi-Source Correlation

Xiatian Zhu
Queen Mary, University of London
xiatian.zhu@qmul.ac.uk

Chen Change Loy
The Chinese University of Hong Kong
ccloy@ie.cuhk.edu.hk

Shaogang Gong
Queen Mary, University of London
sgg@eecs.qmul.ac.uk

1 Introduction

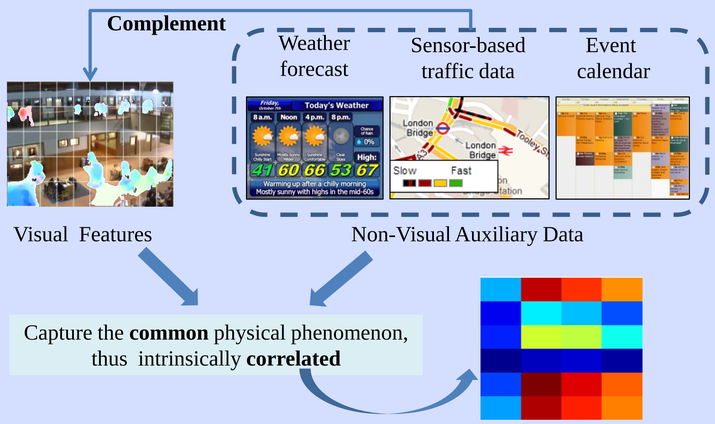


Problem:
How to generate **semantic** synopsis given long video streams by exploiting information beyond low-level visual features?

Existing video synopsis methods:
× typically rely on visual cues alone, this is inherently unreliable
× difficult to bridge the semantic gap between low-level visual features and high-level semantic content interpretation required for better summarisation

Contributions:
✓ Generate semantic video synopsis by jointly learning heterogeneous data sources in an unsupervised manner
✓ Handle missing non-visual data

2 Motivation

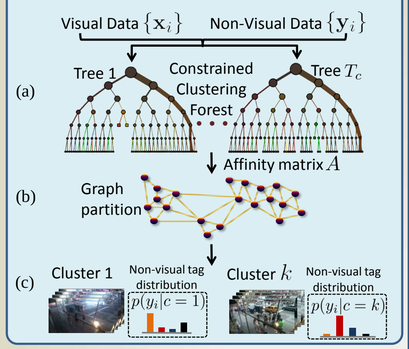


Non-trivial problem that requires joint learning to discover latent associations between heterogeneous multiple data sources:

- Heteroscedasticity problem, e.g. very different representations
- Individual data sources can be inaccurate and incomplete
- Non-visual data is not always available, nor synchronised with visual data

3 Learning a multi-source video synopsis model

Training a synopsis model (overview)



Step (a): Constrained Clustering Forest (CC-Forest)

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

where

$\Delta\mathcal{I}$: the total information gain
$\Delta\mathcal{I}_v,\ \Delta\mathcal{I}_i,\ \Delta\mathcal{I}_t$: gain in individual sources
$\mathcal{I}_{v0},\ \mathcal{I}_{i0},\ \mathcal{I}_{t0}$: inherent source impurity
$\alpha_v, \alpha_i, \alpha_t$: source weights, with $\alpha_v + \sum_{i=1}^{m} \alpha_i + \alpha_t = 1$

Merits of the proposed CC-Forest:

✓ **Joint optimisation** of individual information gain
✓ **Isolate different characteristics** of different sources
✓ **Accommodate partial** or completely **missing** non-visual data

Handle missing non-visual data

An adaptive source weighting method:
1. Reweight the i-th non-visual source as: $\alpha_i - \delta_i \alpha_i$ with δ_i the missing ratio
2. Renormalise all source weights to ensure: $\alpha_v + \sum_{i=1}^{m} \alpha_i + \alpha_t = 1$

Step (b-c): Multi-Source Latent Cluster Discovery

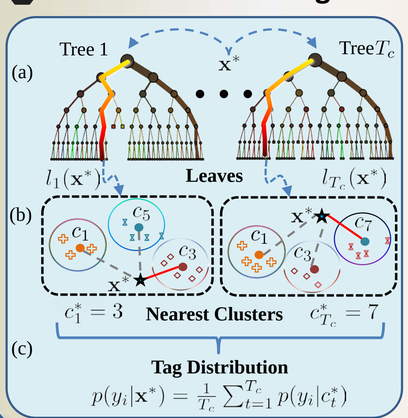
(1) Derive a multi-source-aware affinity matrix A from a learned CC-Forest:
$A = \frac{1}{T_c}\sum_{t=1}^{T_c} A^t$ where A^t is a tree-level affinity, with element defined as:
$A^t_{i,j} = \exp^{-\mathrm{dist}^t(\mathbf{x}_i,\mathbf{x}_j)}$ with $\mathrm{dist}^t(\mathbf{x}_i,\mathbf{x}_j) = \begin{cases} 0 & \text{if } l(\mathbf{x}_i) = l(\mathbf{x}_j), \\ +\infty & \text{otherwise.} \end{cases}$

(2) Symmetrically normalise the affinity matrix, obtain
$S = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ where D denotes a diagonal matrix with elements $D_{i,i} = \sum_j A_{i,j}$

(3) Perform spectral clustering [3] on S, with automatically estimated cluster number
Each training sample $\mathbf{x}_i$ is then assigned to a cluster c_i

(4) Predict a unique distribution of each non-visual data for a cluster c
$p(y_i|c) \propto \sum_{\mathbf{x}_j \in \mathbf{X}_c} p(y_i|\mathbf{x}_j)$ where $\mathbf{X}_c$ refers to the training sample set in c

4 Structure-driven tag inference



Infer non-visual tag of a test sample $\mathbf{x}^*$

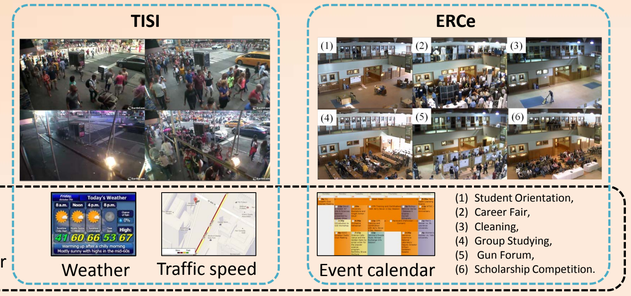
Step (a): trace the target leaf of tree
- search for the leaf of each tree $\mathbf{x}^*$ falls into

Step (b): retrieve leaf level clusters
- derived from training samples sharing the same leaf node
- search for nearest clusters whose tag distribution is used as tree-level prediction

Step (c): average tree-level predictions
- yield a smooth prediction

5 Datasets

Two datasets collected from publicly available webcams: TIme Square Intersection (**TISI**) and Educational Resource Centre (**ERCe**) dataset



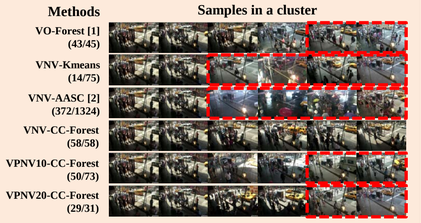
Non-visual auxiliary data:
TISI: weather, traffic speed
ERCe: campus event calendar

6 Clustering evaluation

Table 1. Mean entropy of cluster NV tag distribution **(Red: the best)**

Dataset	TISI		ERCe
Method	traffic speed	weather	event
VO-Forest [1]	0.8675	1.0676	0.0616
VNV-Kmeans	0.9197	1.4994	1.2519
VNV-AASC [2]	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

TISI: cluster purity example – sunny **(Red box: errors)**



* **Our methods;** VO = visual only; VNV = visual + non-visual; VPNV*xx* = *xx*% missing ratio of the training non-visual data.

7 Tag inference evaluation

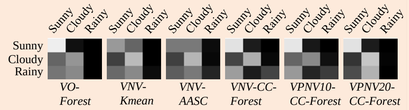
Table 2. TISI: tag inference accuracy comparison **(Red: the best)**

Method	VO-Forest [1]	VNV-Kmeans	VNV-AASC [2]	VNV-CC-Forest	VPNV10-CC-Forest	VPNV20-CC-Forest
traffic speed	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

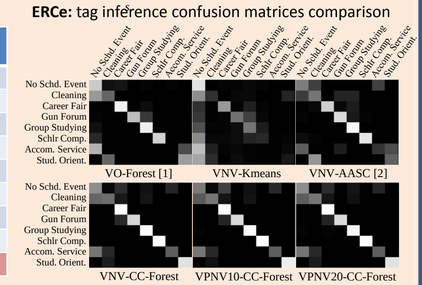
Table 3. ERCe: tag inference accuracy comparison **(Red: the best)**

Method	VO-Forest [1]	VNV-Kmeans	VNV-AASC [2]	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	00.00	00.00	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**

TISI: tag inference confusion matrices comparison



ERCe: tag inference confusion matrices comparison



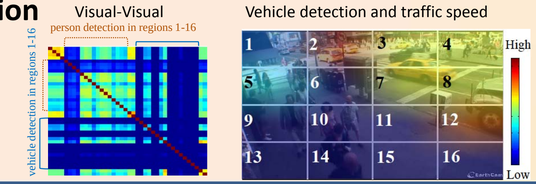
8 Semantic video synopsis

TISI: A synopsis of weather+traffic changes

ERCe: summarisation of some key events

9 Source association

TISI: discovered latent correlations among visual and non-visual sources



[1] L. Breiman. Random forests. ML, 2001
[2] H.-C. Huang, Y.-Y. Chuang, C.-S. Chen. Affinity aggregation for spectral clustering. CVPR, 2012
[3] L. Zelnik-manor and P. Perona. Self-tuning spectral clustering. NIPS, 2004

Project page:
http://www.eecs.qmul.ac.uk/~xz303/