Learning from Multiple Sources for Video Summarisation

Zhu, X; Loy, CC; Gong, S

Noname manuscript No.
(will be inserted by the editor)

Learning from Multiple Sources for Video Summarisation

Xiatian Zhu, · **Chen Change Loy,** · **Shaogang Gong**

Received: date / Accepted: date

Abstract Many visual surveillance tasks, e.g. video summarisation, is conventionally accomplished through analysing imagery-based features. Relying solely on visual cues for public surveillance video understanding is unreliable, since visual observations obtained from public space CCTV video data are often not sufficiently trustworthy and events of interest can be subtle. We believe that non-visual data sources such as weather reports and traffic sensory signals can be exploited to complement visual data for video content analysis and summarisation. In this paper, we present a novel unsupervised framework to learn jointly from both visual and independently-drawn non-visual data sources for discovering meaningful latent structure of surveillance video data. In particular, we investigate ways to cope with discrepant dimension and representation whilst associating these heterogeneous data sources, and derive effective mechanism to tolerate with missing and incomplete data from different sources. We show that the proposed multi-source learning framework not only achieves better video content clustering than state-of-the-art methods, but also is capable of accurately inferring missing non-visual semantics from previously-unseen videos. In addition, a comprehensive user study is conducted to validate the quality of video summarisation generated using the proposed multi-source model.

Xiatian Zhu
School of Electronic Engineering and Computer Science, Queen Mary University of London.
E-mail: xiatian.zhu@qmul.ac.uk

Chen Change Loy
Department of Information Engineering, The Chinese University of Hong Kong.
E-mail: ccloy@ie.cuhk.edu.hk

Shaogang Gong
School of Electronic Engineering and Computer Science, Queen Mary University of London.
E-mail: s.gong@qmul.ac.uk

Keywords Multi-source data · heterogeneous data · visual surveillance · event recognition · video summarisation.

1 Introduction

Visual features and descriptors are often carefully designed and exploited as the sole input for surveillance video content analysis and summarisation. For instance, optical or particle flow is typically employed in activity modelling (Hospedales et al, 2011; Wang et al, 2009; Wu et al, 2010), foreground pixel feature is used for multi-camera video understanding (Loy et al, 2012), space-time image gradient is adopted for crowd analysis (Kratz and Nishino, 2012), and mixture of dynamic textures is used for video segmentation (Chan and Vasconcelos, 2008) and anomaly detection (Li et al, 2013).

A critical task in visual surveillance is to automatically make sense of 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 only information source for video summarisation (Kang et al, 2006; Pritch et al, 2008; Feng et al, 2012; Lee et al, 2012; Lu and Grauman, 2013a). On the other hand, in complex and cluttered public scenes there are intrinsically more interesting and salient higher-level events that can provide more meaningful and concise summarisation of the video data. However, such events may not be visually well-defined (easily detectable) nor 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 (Gong et al, 2011).

[1] Spatio-temporal combinations of human activity or interaction patterns, e.g. gathering, or environmental state changes, e.g. raining.

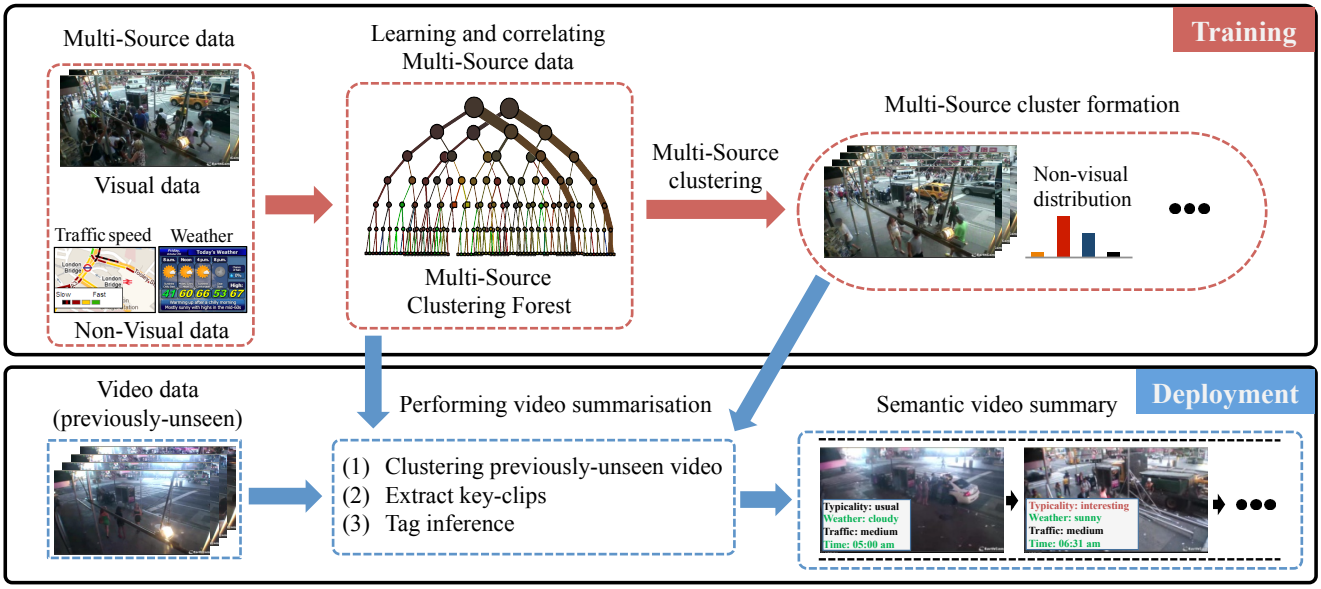


Fig. 1: The overview of the proposed multi-source driven video summarisation framework. We consider a novel setting where multiple heterogeneous sources are present during the model training stage. The proposed Multi-Source Clustering Forest discovers and exploits latent correlations among heterogeneous visual and non-visual data sources both of which can be inaccurate and not trustworthy. In deployment, our model uncovers visual content structures and infer semantic tags on previously-unseen video data for video summarisation.

In this study, we wish to exploit non-visual auxiliary information to complement the unilateral perspective from visual observations. Examples of non-visual sources include weather report, GPS-based traffic data, geo-location data, textual data from social networks, and on-line event schedules. The auxiliary data sources are beneficial to visual data modelling because despite that visual and non-visual data may have very different characteristics and are of different natures, they depict the common physical phenomenon in a scene. They are intrinsically correlated, although may be mostly indirect in some latent spaces. Effectively discovering and exploiting such a latent correlation space can facilitate the underlying data structure discovery and bridge the semantic gap between low-level visual features and high-level semantic interpretation.

Challenges - Nevertheless, it is non-trivial to formulate a framework that exploits both visual and non-visual data for video content analysis and summarisation, both algorithmically and in practice.

Algorithmically, unsupervised mining of latent correlations and interactions between heterogeneous data sources faces a number of challenges: (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 great extent ($>$2000 visual dimensions vs. $<$10 non-visual dimensions). Owing to this dimensionality discrepancy problem, a straightforward concatenation of features will result in a representation unfavourably inclined towards the imagery data. (2) Both visual and non-visual data in isolation can be inaccurate and incomplete.

In practice, auxiliary data sources, e.g. weather, traffic reports, and event time tables, may be rather unreliable in availability. Specifically, the reports may not be released on-the-fly at a synchronised time stamp with the surveillance video stream. In addition, existing video control rooms may not necessarily have direct access to these sources. This renders models that expect complete visual and non-visual information during deployment impractical.

Our solution - In this study, we address this multi-source learning problem in the context of video summarisation, conventionally based on visual feature analysis and object detection or segmentation. In particular, we formulate a novel framework that is capable of performing joint learning given heterogeneous multi-sources (Figure 1). We consider visual data as the *main source* and non-visual data as the *auxiliary sources*, since we believe visual information still plays the main role in video content analysis. During training, we assume the access to both visual and non-visual data. The model performs multi-source data clustering and discovers a set of visual clusters tagged along with non-visual data distribution, e.g. different weathers and traffic speeds. We term the model as *multi-source model*. During the deployment stage, we only assume the availability of previously-unseen

[2] Also known as the heteroscedasticity problem (Duin and Loog, 2004).

video data since non-visual data may not be accessible due to the aforementioned limitations. Since the learned model has already captured the latent structure of heterogeneous types of data sources, the model can be used for semantic video clustering and non-visual tag inference on previously-unseen video sequence, even without the non-visual data. Subsequently, key clips are automatically selected from the discovered clusters. The final summary video can be produced by chronologically compositing these key clips enriched by the inferred tags.

Contributions - The main contributions of this work are:

1. We propose a unified multi-source learning framework capable of discovering semantic structures of video content collectively from heterogeneous visual and non-visual data. This is made possible by formulating a novel Multi-Source Clustering Forest (MSC-Forest) that seamlessly handles multi-heterogeneous data sources dissimilar in representation, distribution, and dimension. 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 better data structure discovery.
2. The model is novel in its ability to accommodate partial or completely missing non-visual sources. In particular, we introduce a joint information gain function that is capable of dynamically adapting to arbitrary amount of missing non-visual information during model learning. In model deployment, only visual input is required for inferring missing non-visual semantics.

Extensive comparative evaluations are conducted on two public surveillance videos captured from both indoor and outdoor environments. Comparative results show that the proposed model not only outperforms the state-of-the-art methods (Huang et al, 2012; Criminisi and Shotton, 2012) for video content clustering and structure discovery, but also is more superior in predicting non-visual tags for previously-unseen videos. The robustness of the proposed model is further validated by a user study on video summary quality.

2 Related Work

Multi-modality learning - There exist studies that exploit different sensory or information modalities from a single source for data structure mining. For example, Cai et al. (Cai et al, 2011) propose to perform multi-modal image clustering by learning a commonly shared graph-Laplacian matrix from different visual feature modalities. Heer and Chi (Heer and Chi, 2001) combine linearly individual similarity matrices derived from multi-modal webpages for web user grouping. Karydis et al. (Karydis et al, 2009) present a tensor based model to cluster music items with additional tags. In terms of video analysis, the auditory channel and/or transcripts have been widely explored for detecting semantic concepts from multimedia videos (Zhang et al, 2004; Fu et al, 2013), summarising highlights in news and broadcast programs (Taskiran et al, 2006; Gong, 2003), or locating speakers (Khalidov et al, 2011). User tags associated with web videos (e.g. YouTube) have also been utilised (Wang et al, 2010; Toderici et al, 2010; Wang et al, 2012). In contrast, surveillance videos captured from public spaces are typically without auditory signals nor any synchronised transcripts and user tags available. Instead, 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.

Multi-source learning - An alternative multi-source learning mechanism can be clustering ensemble (Strehl and Ghosh, 2003; Topchy et al, 2005) where a collection of clustering instances is generated and then aggregated into the final clustering solution. Typically only single data source is considered, but it can be easily extended to handle multi-source data, e.g. creating a respective clustering instance for each source. Nonetheless, cross-source correlation is ignored since the clustering instances are separately formed and no interaction between them is involved. A closer approach to ours is the Affinity Aggregation Spectral Clustering (AASC) (Huang et al, 2012), which learns data structure from multiple types of homogeneous information (visual features only). Their method generates independently multiple affinity data matrices by exhaustive pairwise distance computation for every pair of samples in every data source. It suffers from unwieldy representation given high-dimensional data inputs. Importantly, despite that it seeks for optimal weighted combination of distinct affinity matrices, it does not consider correlation between different sources in model learning, similar to clustering ensemble (Strehl and Ghosh, 2003; Topchy et al, 2005). Differing from the above models, our Multi-Source Clustering Forest overcomes these problems by generating a unified single affinity matrix that captures latent correlations among heterogeneous types of data sources. Furthermore, our model has a unique advantage in handling missing non-visual data over (Strehl and Ghosh, 2003; Topchy et al, 2005; Huang et al, 2012).

Video summarisation - Contemporary video summarisation methods can be broadly classified into three paradigms: (1) key-frame-based (Kim et al, 2014; Khosla et al, 2013; Lee et al, 2012; Cong et al, 2012; Wolf, 1996; Zhang et al, 1997; Truong and Venkatesh, 2007; Money and Agius, 2008), (2) segment-based (Gygli and Van Gool, 2015; Chu et al, 2015; Sun et al, 2014; Potapov et al, 2014; Gygli et al, 2014; Zhao and Xing, 2014; Lu and Grauman, 2013b; Cong et al, 2012), and (3) object-based (Pritch et al, 2008; Feng et al, 2012; Pritch et al, 2007; Lin et al, 2015) methods.

Specifically, the key-frame-based approaches select representative key-frames by analysing low-level imagery properties such as optical flow (Wolf, 1996) or image differences (Zhang et al, 1997), by modelling object's appearance and motion (Lee et al, 2012), or by forming a storyboard of still images through exploiting internet-images using either a single photographer (Kim et al, 2014) or a collection of user-provided images (Khosla et al, 2013). Similarly, the aim of video segment based methods is to identify interesting and representative short moments. Different measurement and selection criteria have been exploited. For example, Chu et al (2015) consider visual co-occurrence among the same-topic videos as a content importance measure; both Zhao and Xing (2014) and Cong et al (2012) treat video summarisation as a sparse coding problem wherein a dictionary based reconstruction error is used as the selection standard; Gygli et al (2014) consider the interestingness of visual content; Potapov et al (2014) and Sun et al (2014) assume that video category are known *a priori* and measure shot importance with category-specific learned models, such as SVMs; Gygli and Van Gool (2015) perform joint learning of multiple objectives (e.g. representativeness, interestingness and uniformity) over training summary videos for extracting global importance from raw videos. Lu and Grauman (2013b) focus on the connectivity and coherency of the generated summary storyline by selecting video parts with important shot-to-shot influence. Object-based summarisation techniques (Pritch et al, 2008; Feng et al, 2012; Lin et al, 2015), on the other hand, rely on object segmentation and tracking to extract object-centric trajectories/tubes, and compress those tubes to reduce spatio-temporal redundancy. In particular, Pritch et al (2008) and Feng et al (2012) summarising all detected motion trajectories, whilst Lin et al (2015) additionally consider the abnormality and category nature of individual motions by modelling localised spatio-temporal blobs and composite category-specific summary videos using only abnormal object-tubes.

All the above schemes utilise solely visual information and make implicit assumptions about the completeness and accuracy of the visual data available in extracting visual features or object-centered representations. They are unsuitable nor scalable to complex scenes where visual data are inherently incomplete and inaccurate, mostly the case in surveillance videos. Our work differs significantly to these studies in that we exploit not only visual data without object tracking, but also non-visual sources as complementary information. The summary generated by our approach is semantically enriched – it is labelled automatically with semantic tags, e.g. traffic condition, weather, or event. All these tags are learned from heterogeneous non-visual sources in an unsupervised manner during model training without any manual labels.

Random forests - Random forests (Breiman, 2001; Criminisi and Shotton, 2012) have proven as powerful models in the literature. Different variants of random forests have been devised, either supervised (Shotton et al, 2011; Gall et al, 2011; Schulter et al, 2013a; Bosch et al, 2007; Caruana et al, 2008), or unsupervised (Liu et al, 2000; Shi and Horvath, 2006; Perbet et al, 2009; Moosmann et al, 2008; Zhu et al, 2013, 2014, 2015). Supervised models are not suitable to our problem since we do not assume the availability of ground truth labels during model training. Existing clustering forest models, on the other hand, assumes only homogeneous data sources such as pure imagery-based features. No principled way of combining multiple heterogeneous and independent data sources in forest models is available.

3 Multi-Source Clustering

Video summarisation by content abstraction aims to generate a compact summary composed of key/interesting content from a long previously-unseen video for achieving efficient holistic understanding (Truong and Venkatesh, 2007). A common way to establish a video summary is by extracting and then combining a set of key frames or shots. These key contents are usually discovered and selected from clusters of video frames or clips (Truong and Venkatesh, 2007).

In this study, we follow the aforementioned approach but consider not only visual content of video, but also a large corpus of non-visual data collected from heterogeneous independent sources (Figure 2(a)). Specifically, through learning latent structure of multi-source data (Figure 2(b-c)), we wish to make reference to and/or impose non-visual semantics directly into video clustering without any human manual annotation of video data (Figure 2(d)). Formally, we consider the following different data sources that form a multi-source input feature space:

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

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 $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,m}) \in \mathbb{R}^m, i = 1, \ldots, n$. Note that any (or all) dimension of $\mathbf{y}_i$ may be missing.

We aim at formulating a unified clustering model capable of coping with the few challenges as highlighted in Section 1. The model needs be unsupervised since no ground truth is assumed. To mitigate the heteroscedasticity and dimension discrepancy problems, we require a model that can isolate the very different characteristics of visual and non-visual data, yet can still exploit their latent correlation in the

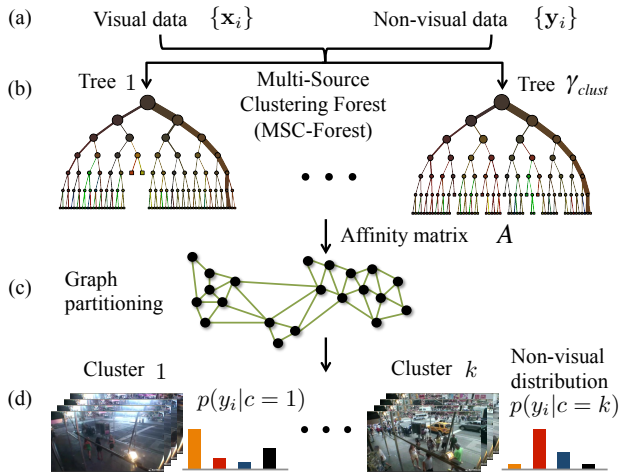


Fig. 2: Multi-source model training stage: The pipeline of performing multi-source clustering on visual and non-visual data with the proposed Multi-Source Clustering Forest (MSC-Forest).

clustering process. To handle noisy data, feature selection is needed and necessary.

In light of the above demands, we choose to start with the clustering random forest (Breiman, 2001; Liu et al, 2000; Shi and Horvath, 2006) due to (1) unsupervised information gain optimisation thus requiring no ground truth labels; (2) its flexible objective function for facilitating the modelling of multi-source data as well as the processing of missing data; (3) and its implicit feature selection mechanism for handling noisy features. Nevertheless, the conventional clustering forest is not well suited to solve these challenges since it expects a full concatenated representation as input during both model training and deployment. This does not conform to the assumption of only visual data being available during model deployment for previously-unseen videos. Moreover, due to its uniform variable selection mechanism (Breiman, 2001) (e.g. each feature dimension has the same probability to be selected as a candidate optimal splitting variable), there is no principled way to ensure balanced contribution from individual visual and non-visual sources in the node splitting process. To overcome these limitations, we propose a new *Multi-Source Clustering Forest* (MSC-Forest) by introducing a new objective function allowing *joint optimisation of individual information gains* of different sources. We first describe the conventional forests prior to detailing the proposed MSC-Forest.

3.1 Conventional Random Forests

Classification forests - A general form of random forests is the classification forests. A classification forest (Breiman, 2001; Schulter et al, 2013b) is an ensemble of γ_{class} binary decision trees $f_{\text{tree}}(\mathbf{x})$: $\mathring{X} \rightarrow \mathbb{R}^{\mathring{k}}$, with $\mathring{X}$ the d-dimensional feature space, and $\mathbb{R}^{\mathring{k}} = [0,1]^{\mathring{k}}$ denoting the space of class probability distribution over the label space $\mathring{L} = \{1, \ldots, \mathring{k}\}$.

Decision trees are learned independently of each other, each with a random subset X_t of the training samples $X = \{\mathbf{x}_i\}$, i.e. bagging (Breiman, 2001). Growing a decision tree involves a recursive node splitting procedure until some stopping criterion is satisfied, e.g. leaf nodes are formed when no further split can be achieved given the objective function, or the number of training samples arriving at a node is smaller than the predefined node size, ϕ. Small ϕ leads to deep trees. We set $\phi = 2$ in our experiments for capturing sufficiently fine-grained data structure. At each leaf node, the class probability distribution is then estimated based on the labels of the arrival samples.

The training of each internal/split node is a process of binary split function optimisation, defined as

$$\mathrm{h}(\mathbf{x}, \mathbf{w}) = \begin{cases} 0 & \text{if } x_\kappa < \vartheta, \\ 1 & \text{otherwise.} \end{cases} \tag{1}$$

This split function is parameterised by two parameters $\mathbf{w} = [\kappa, \vartheta]$: (i) a feature dimension x_κ with $\kappa \in \{1, \ldots, d\}$, and (ii) a feature threshold $\vartheta \in \mathbb{R}$. All samples of a split node s will be channelled to either the left l or right r child nodes, according to the output of Eqn. (1).

The optimal split parameter $\mathbf{w}^*$ is chosen via

$$\mathbf{w}^* = \underset{W}{\operatorname{argmax}}\, \Delta\psi_{\text{class}}, \tag{2}$$

where $W = \{\mathbf{w}^i\}_{i=1}^{m_{\text{try}}(|S|-1)}$ represents a parameter set over m_{try} randomly selected features, with S the sample set reaching the node s. The cardinality of a set is given by $|\cdot|$. Typically, a greedy search strategy is exploited to identify $\mathbf{w}^*$. The information gain $\Delta\psi_{\text{class}}$ is formulated as

$$\Delta\psi_{\text{class}} = \psi_s - \frac{|L|}{|S|}\psi_l - \frac{|R|}{|S|}\psi_r, \tag{3}$$

where L and R denote the sets of data routed into l and r, and $L \cup R = S$. The information gain ψ can be computed as either the entropy or Gini impurity (Breiman et al, 1984).

Clustering forests - In contrast to classification forests, clustering forests require no ground truth label information during the training phase. A clustering forest consists of γ_{clust} binary decision trees. The leaf nodes in each tree define a spatial partitioning of the training data. Interestingly, the training of a clustering forest can be performed using the classification forest optimisation approach by adopting the pseudo two-class algorithm (Breiman, 2001; Liu et al, 2000; Shi and Horvath, 2006). Specifically, we add N pseudo samples $\bar{\mathbf{x}} = \{\bar{x}_1, \ldots, \bar{x}_d\}$ (Figure 3(b)) into the original data space X (Figure 3(a)), with $\bar{x}_i \sim \text{Dist}(x_i)$ sampled from certain distributions $\text{Dist}(x_i)$. In the proposed model, we

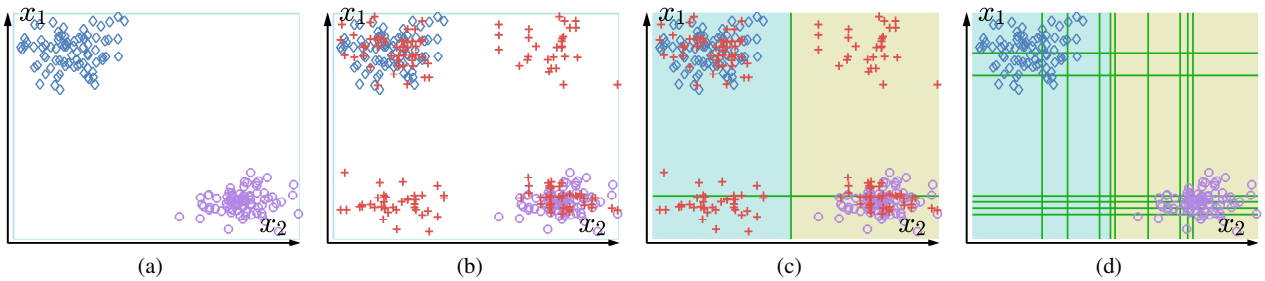


Fig. 3: An illustration of clustering toy data with a clustering forest. (a) Original toy data are labelled as class 1, whilst (b) the pseudo-points (red +) as class 2. (c) A clustering forest performs two-class classification in the augmented space. (d) The final data partitions on the original data.

adopt the empirical marginal distributions of the feature variables owing to its favourable performance (Shi and Horvath, 2006). With this data augmentation strategy, the clustering problem becomes a canonical classification problem that can be solved by the classification forest training method as discussed above. The key idea behind this algorithm is to partition the augmented data space into dense and sparse regions (Figure 3(c-d)) (Liu et al, 2000).

3.2 Multi-Source Clustering Forest

Conventional clustering forests assumes only homogeneous data sources such as pure imagery-based features. In contrast, the proposed Multi-Source Clustering Forest can take heterogeneous sources as input. In particular, the proposed model uses visual features as splitting variables to grow Multi-Source Clustering trees (MSC-trees) as in Eqn. (1), and exploits non-visual information as additional data to help determining the $\mathbf{w} = [\kappa, \vartheta]$. In this way, auxiliary non-visual information is used, in addition to visual data, to guide the tree formation.

Formally, we define a new joint information gain function for node splitting during training MSC-trees as:

$$\Delta\psi = \underbrace{\alpha_v \frac{\Delta\psi_v}{\psi_{v0}}}_{\text{visual}} + \underbrace{\sum_{j=1}^{m} \alpha_j \frac{\Delta\psi_j}{\psi_{j0}}}_{\text{non-visual}} + \underbrace{\alpha_t \frac{\Delta\psi_t}{\psi_{t0}}}_{\text{temporal}}. \quad (4)$$

Similar to Eqn. (3), the optimal parameter corresponds to the split with the maximal $\Delta\psi$. This formulation defines the best data split across the joint space of multi-source data, beyond visual domain alone. All the terms in Eqn. (4) are interpreted as below.

Visual term: $\Delta\psi_v = \Delta\psi_{\text{class}}$ (Eqn. (3)) denotes the information gain in visual domain. Precisely, this measure is computed from the pseudo class labels. Therefore, it reflects the visual data structure characteristics given that the pseudo data samples are drawn from the marginal feature distributions (Section 3.1). In this study we utilise the Gini impurity e_{gini} (Breiman et al, 1984) to estimate $\Delta\psi_{\text{class}}$ by setting $\psi = e_{\text{gini}}$ in Eqn. (3) due to its simplicity and efficiency. The Gini impurity is computed as $e_{\text{gini}} = \sum_{i \neq j} p_i p_j$, with p_i and p_j being the proportion of samples belonging to the ith and jth category in a split node s. High value in e_{gini} indicates pure category distribution.

Non-visual term: This is a new term we introduce as auxiliary information on visual term. More specifically, $\Delta\psi_j$ denotes the information gain in the jth non-visual data. A non-visual source can be either categorical or continuous. For a categorical non-visual source, similar to visual term we use the Gini impurity e_{gini} as its data split measure criterion. In the case of non-visual source with continuous values, we adopt least squares regression (Breiman et al, 1984) to enforce continuity in the clustering space:

$$e_{\text{lsr}} = \frac{1}{|S|} \sum_{i=1}^{|S|} (y_{i,j} - \frac{1}{|S|} \sum_{i=1}^{|S|} y_{i,j})^2, \quad (5)$$

where $y_{i,j}$ represents the value in the jth non-visual space associated with the ith sample $\mathbf{x}_i \in S$, and S is the set of samples reaching node s. That is $\Delta\psi_j = e_{\text{lsr}}$.

Temporal term: We add a temporal smoothness gain $\Delta\psi_t$ to encourage temporally adjacent video clips to be grouped together. The intuition is that human activity/event patterns may present a great deal of disparity at different times of a day, e.g. day *versus* night, or morning *versus* afternoon. In other words, activity video semantic structure is inherently time-dependent. Therefore, this temporal information can help in mining visual data structure. Specifically, we utilise the video recording time associated with video clips as a temporal-constraint, and exploited the least squares regression (Eqn.(5)) to compute its information gain since time is continuous.

The information gain by different sources may live in very disparate ranges due to the different natures of source, each term of Eqn. (4) is therefore normalised by its initial data impurity denoted by ψ_{v0}, ψ_{j0}, and ψ_{t0}. These impurities are obtained at the root node of every MSC-tree. The

source weights are denoted by α_v, α_i, and α_t accordingly, holding $\alpha_v + \sum_{i=1}^{m} \alpha_i + \alpha_t = 1$. We set $\alpha_v = 0.5$ obtained by cross-validation. A detailed analysis on α_v is given in Section 6.2. For non-visual and temporal information, we uniformly assign $\alpha_t = \alpha_i = \frac{1-\alpha_v}{m+1}$ since their importance is not known *a prior*, with m the number of non-visual sources.

The role of different source data - Given the main role and much more stable provision of the visual source in video understanding, non-visual data are regarded as auxiliary information over visual source. During the training of MSC-Forest, the split functions (Eqn. (1)) are defined on visual features, but $\mathbf{w} = [\kappa, \vartheta]$ is collectively determined by visual features and the associated non-visual as well as temporal information (i.e. the non-visual and temporal term in Eqn. (4)). Alternatively, one can think of that the *main* visual data source is 'completely-visible' to the MSC-Forest since it is needed during both forest training and evaluation, whilst the *auxiliary* non-visual data are 'half-visible' in that they are exploited as side information for embedding their knowledge into the MSC-tree growing during model training but not required any more during the MSC-Forest evaluation (due to their restricted availability as explained in Section 1).

Joint information gain - We interpret the intrinsic advantage of the joint information gain defined by Eqn. (4), with comparison against the naïve feature concatenation strategy. With the latter scheme, the information gain (Eqn. (3)) is directly estimated in a heterogeneous joint space where visual, non-visual and temporal data are mixed together. This would suffer from the heteroscedasticity problem, as discussed in Section 1. Instead, Eqn. (4) overcomes this challenge by modelling different sources via separate information gain terms, resulting in a more balanced exploitation of multi-source data. In this way, the proposed joint information gain of multi-source data encourages more appropriate visual data separation both visually and semantically. This formulation is the essential contribution of our proposed MSC-Forest model.

The merits of MSC-Forest - The formulation in Eqn. (4) brings two unique benefits: (A) Thanks to the information gain optimisation, the influences of visual and non-visual domains on data partitioning can be better balanced compared to naïve feature concatenation. (B) Eqn. (2) and Eqn. (4) together provide a mechanism to discover strongly correlated heterogeneous source pairs and to exploit joint information gain of such correlated pairs for data partitioning. In other words, only selective visual features (Eqn. (2)) that yield high information gain collectively with non-visual information (Eqn. (4)) will contribute to the MSC-tree growing. Such a mechanism cannot be realised using the conventional clustering forests (Breiman, 2001; Liu et al, 2000). We shall demonstrate the multi-source correlation discovered by our proposed MSC-Forest in experiments (Section 6.4).

3.2.1 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[3]. Specifically, when some non-visual data are missing and suppose the missing proportion of the ith non-visual type in the training set X_t for MSC-tree t is δ_i, we reduce its weight from α_i to $\alpha_i - \delta_i\alpha_i$. The total reduced weight $\sum_i \delta_i\alpha_i$ is then distributed evenly to the weights of all sources to ensure $\alpha_v + \sum_{i=1}^{m} \alpha_i + \alpha_t = 1$. This linear adaptive weighting method produces satisfactory results in our experiments.

3.2.2 Model Complexity

The upper-bound learning complexity of a whole MSC-Forest can be examined from its constituent parts, i.e. at tree- and node-levels. Formally, given a MSC-tree t, we denote the set of all the split nodes as Π_t and the sample subset used for training a split node $j \in \Pi_t$ as S_j. The training complexity of j-th node is given by $m_{\text{try}}(|S_j| - 1)u$, when a greedy search algorithm is adopted, with m_{try} the number of features attempted to partition S_j, and u the running time of conducting one data splitting operation. Consequently, the overall computational cost of learning a MSC-Forest can be computed as

$$\begin{aligned} l_{\text{cost}} &= \sum_{t}^{\gamma_{\text{clust}}} \sum_{j \in \Pi_t} m_{\text{try}}(|S_j| - 1)u \\ &= m_{\text{try}} u \sum_{t}^{\gamma_{\text{clust}}} \sum_{j \in \Pi_t} (|S_j| - 1). \end{aligned} \tag{6}$$

The value of parameter m_{try} is identical across all MSC-trees. The learning time is thus determined by (1) the value of u, and (2) the factor that we name as *tree fan-in*

$$\varpi(t) = \sum_{j \in \Pi_t} |S_j - 1|. \tag{7}$$

Clearly, u of a MSC-Forest is larger than that of conventional forests since we need to compute additional information gains of non-visual and temporal information (Eqn. (4)). On the other hand, the value of $\varpi(t)$ primarily relies on

[3] There exist missing data filling algorithms utilised in conventional random forests, e.g. for the missing value of one feature in one class, the median value (continuous) or the most frequent category (discrete) of this feature over the current class can be used as the estimation (Breiman, 2003). Whilst a similar strategy is possible to apply on our MSC-Forest, we consider an alternative by proposing an effective adaptive weighting algorithm in order not to further introduce noisy training data.

the tree structure/topological characteristics (Martin, 1997): a balanced and shallower tree has smaller $\varpi(t)$, thus the tree shall be more efficient in training and inference on previously-unseen samples, in that the paths from the root to leaf nodes are relatively shorter. In Section 6.5, we will show that the additional non-visual information encourages more balanced and shallower decision trees than learning from single visual source alone.

3.3 Latent Multi-Source Data Structure Discovery

Given heterogeneous feature spaces involving visual and non-visual data, it is non-trivial to discover their underlying group structures, due to the heteroscedasticity problem aforementioned (Section 1). To this end, MSC-Forest is particularly designed to principally extract and combine the information from multiple individual sources so as to more accurately measure data pairwise similarity relations, which in turn facilitates existing graph-based clustering algorithm, e.g. spectral clustering, to eventually reveal the latent data clusters. Figure 2 depicts the pipeline of our video data clustering approach based on the learned MSC-Forest.

The spectral clustering (Zelnik-manor and Perona, 2004) groups data using eigenvectors of an affinity matrix derived from the data. The goodness of the resulting cluster formation primarily relies on the quality of the input affinity matrix which reflects and embeds the essential data structures (Zhu et al, 2014). Below we describe the details of constructing multi-source referenced affinity matrix from MSC-Forest. Intuitively, the multi-source learning nature of MSC-Forest renders its data similarity measure sensitive to the joint knowledge from diverse source data.

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

$$\mathrm{dist}(\mathbf{x}_i,\mathbf{x}_j) = \begin{cases} 0 & \text{if } \ell(\mathbf{x}_i) = \ell(\mathbf{x}_j), \\ +\infty & \text{otherwise.} \end{cases} \tag{8}$$

We assign the maximum affinity (affinity=1, distance=0) between points $\mathbf{x}_i$ and $\mathbf{x}_j$ if they fall into the same leaf, and the minimum affinity (affinity=0, distance=1) otherwise. A smooth affinity matrix can be obtained through averaging all the tree-level affinity matrices

$$\mathbf{A} = \frac{1}{\gamma_{\mathrm{clust}}} \sum_{t=1}^{\gamma_{\mathrm{clust}}} \mathbf{A}^t, \tag{9}$$

Eqn. (9) is adopted as the ensemble model of MSC-Forest due to its advantage of suppressing the noisy tree predictions, though other alternatives such as the product of tree-level predictions are possible (Criminisi and Shotton, 2012). We then construct a sparse k-NN graph, whose edge weights are defined by the affinity matrix $\mathbf{A}$ (Figure 2(c)).

Subsequently, we symmetrically normalise $\mathbf{A}$ to obtain $\mathbf{S} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$, where $\mathbf{D}$ denotes a diagonal degree matrix with elements $\mathbf{D}_{i,i} = \sum_j^n \mathbf{A}_{i,j}$. Given $\mathbf{S}$, we perform spectral clustering to discover the latent clusters of training clips with the number of clusters automatically determined through analysing the eigenvector structure (Zelnik-manor and Perona, 2004). Each training clip $\mathbf{x}_i$ is then assigned to a cluster $c_i \in C$, with C the set of all clusters.

The learned clusters group similar clips both visually and semantically, with each of the clusters associated with a unique distribution for each non-visual data (Figure 2(d)). We denote the distribution of the ith non-visual data type of the cluster c as

$$p(y_i|c) \propto \sum\nolimits_{\mathbf{x}_j \in X_c} p(y_i|\mathbf{x}_j), \tag{10}$$

where X_c represents the set of training samples in c. These multi-source data clusters form a component of our multi-source model (Figure 1).

4 Semantic Video Summarisation

In Section 3 we presented multi-source data clustering by learning a Multi-Source Clustering Forest (MSC-Forest), resulting in a consistent cluster formation. Once this multi-source model is learned, it can be deployed for semantic video summarisation. Specifically, we follow the established approach of summarising videos by clustering (Truong and Venkatesh, 2007) but with the introduction of two noticeable differences in our method.

Firstly, our *video summary is multi-source referenced*. Specifically, the MSC-Forest is trained on heterogeneous sources, its optimised split functions $\{\mathrm{h}\}$ (Eqn. (1)) therefore implicitly capture the complex multi-source structures. When one deploys the trained model for content summarisation of previously-unseen video data, the model only needs to take visual inputs without any non-visual data sources. And yet it is able to induce video content partitions that not only correspond to visual feature similarities, but also are consistent with meaningful non-visual semantic interpretations. Secondly, our *video summary is automatically tagged* as the result of model inference. This is made possible through exploiting the non-visual data distributions associated with the discovered clusters on the training data (see Eqn. (10) and Figure 2(d)). Below we discuss the details of generating a semantic video summary.

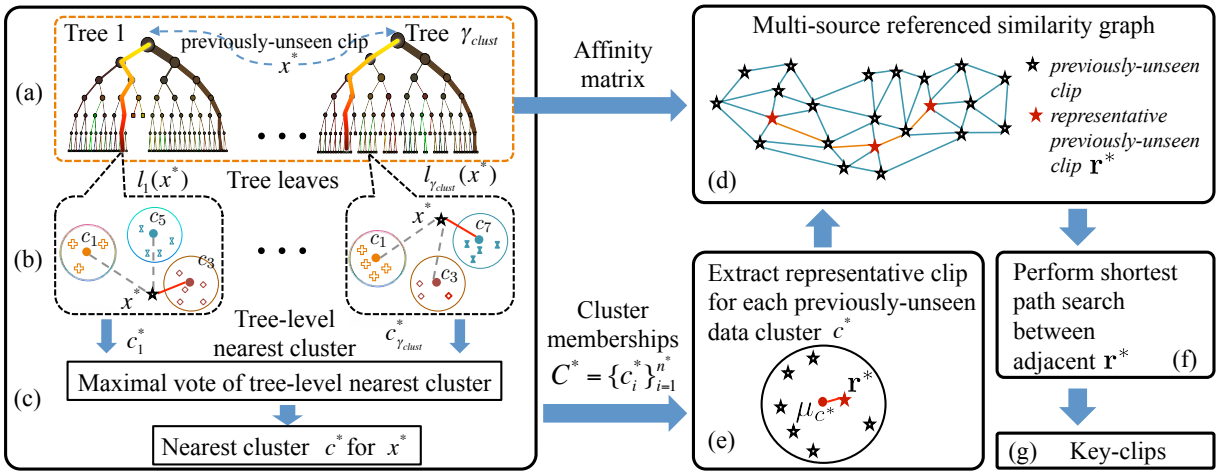


Fig. 4: The pipeline of our multi-source referenced key-clips detection algorithm. (a) Channel a clip $\mathbf{x}^*$ into MSC-trees. (b) Search tree-level nearest clusters of $\mathbf{x}^*$, hollow circle denotes cluster. (c) Predict the final nearest cluster. A red $\star$ depicts a representative previously-unseen clip.

4.1 Key-Clip Extraction and Composition

Suppose we are given a previously-unseen surveillance video footage without meta-data tagging/script. The video is pre-processed by segmenting it into a set of n^* either overlapping or non-overlapping short clips $\{\mathbf{x}_i^*\}_{i=1}^{n^*}$ with equal duration. Our aim is to first assign cluster membership to each previously-unseen clip using the trained multi-source model, and then select key-clips from the resulting clusters[4]. The chosen key-clips are then chronologically ordered to construct a video summary.

Clustering previously-unseen video clips - Inferring cluster memberships of previously-unseen clips is an intricate task. A straightforward method is to assign cluster membership by identifying the nearest cluster $c^* \in C$ to a sample $\mathbf{x}^*$, where C represents the set of clusters we discovered in Section 3.3. However, we found this hard cluster assignment strategy susceptible to outliers in C and source noise. To mitigate this problem, we consider an alternative approach by utilising the MSC-Forest tree structures for soft cluster assignment. This is more robust to either source noise or outliers.

Figure 4 depicts the soft cluster assignment pipeline. First, we trace the leaf $\ell_t(\mathbf{x}^*)$ of each tree t where $\mathbf{x}^*$ falls by channelling $\mathbf{x}^*$ into the tree (Figure 4(a)). This step is critical as it establishes a connection for $\mathbf{x}^*$ with an appropriate training subset $X_{\ell_t(\mathbf{x}^*)}$ using the split functions $\{\mathrm{h}\}_t$ optimised by multi-source data. Here, $X_{\ell_t(\mathbf{x}^*)}$ represents the set of training samples associated with $\ell_t(\mathbf{x}^*)$. The set is consistent with $\mathbf{x}^*$ both visually and semantically since they encompass identical response w.r.t $\{\mathrm{h}\}_t$.

Second, we retrieve the cluster membership $C_t = \{c_i\} \subset C$ of $X_{\ell_t(\mathbf{x}^*)}$, against which we search for the tree-level nearest cluster c_t^* for $\mathbf{x}^*$ (Figure 4(b)) via

$$c_t^* = \mathrm{argmin}_{c \in C_t} ||\mathbf{x}^* - \mu_c||, \tag{11}$$

with t the tree index, and μ_c the centroid of cluster c, estimated as

$$\mu_c = \frac{1}{|X_c|} \sum_{\mathbf{x}_i \in X_c} \mathbf{x}_i, \tag{12}$$

where X_c represents the set of training samples in c. Performing nearest cluster search within C_t rather than the whole cluster space C brings a key benefit: since the search space is constrained by MSC-tree, it is more meaningful and also less noisy than the entire space C, leading to more accurate c_t^* estimation.

Once we obtain all tree-level nearest clusters from all the trees in the forest, $\{c_t^*\}_{t=1}^{\gamma_{\mathrm{clust}}}$, the final nearest cluster c^* is obtained as the one with maximal votes from all the trees (Figure 4(c))

$$c^* = \max \{c_t^*\}_{t=1}^{\gamma_{\mathrm{clust}}} \tag{13}$$

By repeating the above steps on all previously-unseen clips $\{\mathbf{x}_i^*\}_{i=1}^{n^*}$, we obtain their cluster labels as $C^* = \{c_i^*\}_{i=1}^{n^*}$ (Figure 4(e)).

Extracting key-clips - With the assigned cluster memberships C^* on all previously-unseen clips, the key-clip of a previously-unseen video data cluster c^* can be represented

[4] It is worth noticing that the purpose of this clustering step is completely different from the multi-source data clustering during model training, as presented in Section 3.3. The latter is a component of our multi-source model training pipeline (Figure 2), whilst the former aims at revealing the latent structure over testing data for video summarisation.

Algorithm 1: Infer non-visual tags of previously-unseen clips.

Input: A previously-unseen clip $\mathbf{x}^*$, a trained MSC-Forest, training data clusters C;
Output: Predicted tag $\hat{y}_i$;
Initialisation:
  Compute $p(y_i|c)$ for each training data cluster (Eqn. (10));
  Compute cluster centroid μ_c (Eqn. (12));
Non-Visual Tag Inference:
for $t \leftarrow 1$ **to** γ_{clust} **do**
  Trace the leaf $\ell_t(\mathbf{x}^*)$ where $\mathbf{x}^*$ falls (Figure 4(a));
  Retrieve the training samples $X_{\ell_t(\mathbf{x}^*)}$ associated with $\ell_t(\mathbf{x}^*)$;
  Obtain the clusters $C_t = \{c_i\} \subset C$ of $X_{\ell_t(\mathbf{x}^*)}$;
  Search the tree-level nearest cluster c_t^* of $\mathbf{x}^*$ within C_t (Eqn. (11));
end
Estimate tag distribution $p(y_i|\mathbf{x}^*)$ (Eqn. (14));
Compute the final tag $\hat{y}_i$ (Eqn. (15)).

by the representative previously-unseen clip $\mathbf{r}^*$ that is closest to the cluster centroid μ_{c^*} (Figure 4(e)). Concatenating these key-clips chronologically establishes a visual summary. Such a summary, however, is likely to be discontinuous in preserving visual context therefore non-smooth visually due to abrupt changes between adjacent key-clips. To enforce some degrees of smoothness in the visualisation of video summary whilst minimising redundancy, we adopt a shortest path strategy (Boccaletti et al, 2006) to induce an optimal path between two temporally-adjacent representative $\mathbf{r}^*$ on a graph G. This approach produces a visually more coherent video summary whilst discards as much redundancy as possible.

More precisely, we construct a graph $G = (V, E)$, where V and E indicate the set of previously-unseen video clip vertices and edges (Figure 4(d)). The weights of edges can be efficiently estimated using Eqn. (8) and (9). Note that the graph G is also multi-source referenced since it is derived from our multi-source MSC-Forest model. We then perform shortest path search between temporally-adjacent $\mathbf{r}^*$ on G (Figure 4(f)) and all the samples that lie on the shortest paths compose the final key-clip set K_s (Figure 4(g)).

4.2 Video Tagging

Summarising video with high-level interpretation requires plausible semantic content inference from video data $\mathbf{x}^*$. We derive a tree-structure aware tag inference algorithm capable of predicting tag types same as training non-visual data, based on the learned MSC-Forest and discovered training data clusters. Specifically, we first obtain the tree-level nearest cluster c_t^* of a previously-unseen sample $\mathbf{x}^*$ using Eqn. (11). Second, the $p(y_i|c_t^*)$ associated with c_t^* is utilised as the tree-level non-visual tag estimation for the ith non-visual data type. To achieve a smooth prediction, we average all $p(y_i|c = c_t^*)$ obtained from individual trees as

$$p(y_i|\mathbf{x}^*) = \frac{1}{\gamma_{\text{clust}}} \sum_{t=1}^{\gamma_{\text{clust}}} p(y_i|c_t^*). \quad (14)$$

The final tag $\hat{y}_i$ for the ith non-visual type is obtained as

$$\hat{y}_i = \text{argmax}_{y_i}\, p(y_i|\mathbf{x}^*). \quad (15)$$

With the above steps, we can estimate all m non-visual tags $\hat{y}_i$s with $i \in \{1, \ldots, m\}$. The procedure of our tagging algorithm is summarised in Algorithm 1.

Given the extracted key-clips K_s and automatic assignment of non-visual semantic tags (Eqn. (15)), we can now construct a video summary by chronologically concatenating each clip $\mathbf{x}^* \in K_s$ with smooth inter-clip transition, e.g. crossfading, and labelling each clip with their inferred semantic tags.

5 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, and (2) the Educational Resource Centre (ERCe) dataset[5]. 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 Figure 5.

The TISI dataset is challenging due to severe inter-object occlusion, complex behaviour patterns, and large illumination variations caused by both natural and artificial lighting 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.

Table 1: Details of datasets. FPS = frames per second. Training Size = video clip numbers used for model training. Deployment Size = video clip numbers in model deployment.

-	Resolution	FPS	Training Size	Deployment Size
TISI	550×960	10	5819	1505
ERCe	480×640	5	9387	4430

Visual and non-visual sources - We extracted the following set of visual features for representing visual content in each clip: (a) colour features including RGB and HSV; (b) local

[5] Datasets available: www.eecs.qmul.ac.uk/%7Exz303/download.html

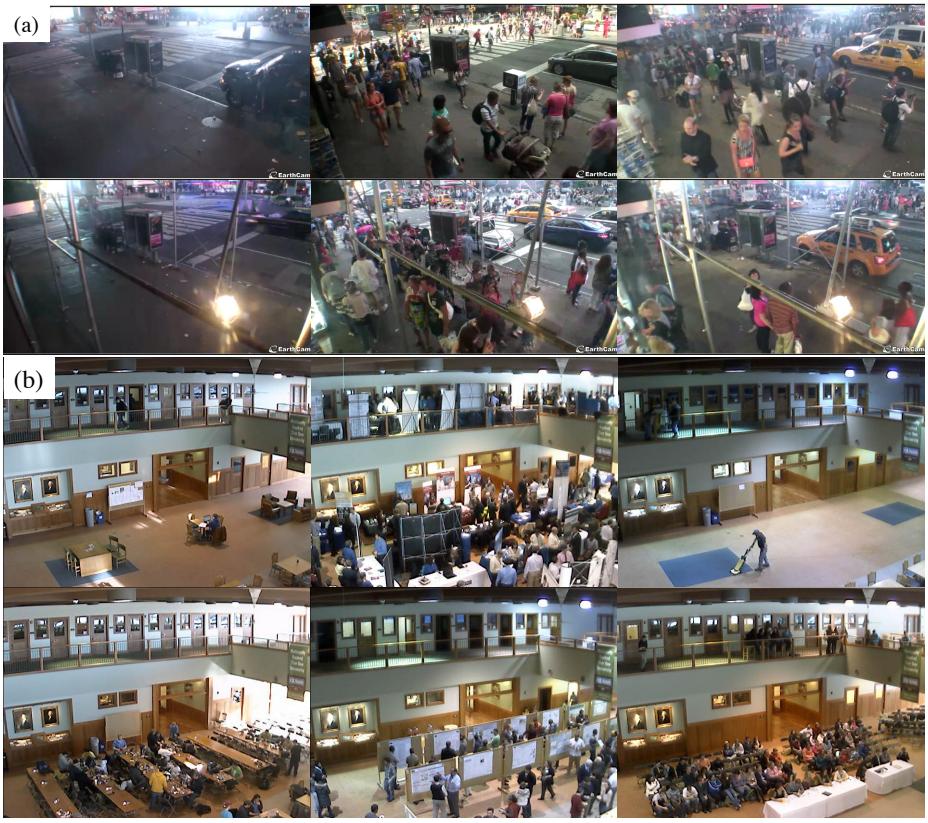


Fig. 5: Examples of the (a) TISI and (b) ERCe datasets.

texture features based on Local Binary Pattern (LBP) (Ojala et al, 2002); (c) optical flow; (d) holistic features of the scene based on GIST (Oliva and Torralba, 2001); and (e) person and vehicle[6] detection (Felzenszwalb et al, 2010).

We collected 10 types of non-visual sources for the TISI dataset: (a) weather data extracted from the WorldWeatherOnline with 9 elements: temperature, weather type, wind speed, wind direction, precipitation, humidity, visibility, pressure, and cloud cover; (b) traffic speed 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 campus events including: No Scheduled Event (NoEvt), Cleaning (Cln), Career Fair (CrF), Gun Forum Control and Gun Violence (GunFrm), Group Studying (GrStd), Scholarship Competition (SchlCpt), Accommodative Service (AcmSvc), Student Orientation (StdOrt).

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

Baselines - To evaluate the proposed method for multi-source video clustering and tag inference, we compared the Visual + Non-Visual + MSC-Forest (*VNV-MSC-Forest*) model against the following baseline models:

1. *VO-Forest*: a conventional forest (Breiman, 2001) trained with visual feature vectors alone, to demonstrate the benefits from using non-visual sources[7].
2. *VNV-Kmeans*: k-means (Jain, 2010) using concatenated vectors of visual and non-visual features, to highlight the heteroscedasticity and dimensionality discrepancy problem caused by heterogeneous visual and non-visual data.
3. *VNV-Forest*: a conventional forest (Breiman, 2001) trained with concatenated visual and non-visual feature vectors, to compare the effectiveness of MSC-Forest that exploits non-visual data during forest formation.
4. *VNV-AASC*: a state-of-the-art multi-source spectral clustering method (Huang et al, 2012) learned by treating each type of visual or non-visual feature as an individual source, to demonstrate the superiority of MSC-Forest in handling diverse data representations and correlating multiple sources.
5. *VNV-COP-Mahal*: a state-of-the-art Mahalanobis distance metric learning method (Xing et al, 2002) using both data features and two types of pairwise constraints, i.e.

[6] No vehicle detection on the ERCe dataset.

[7] Evaluating a forest that takes only non-visual inputs is not possible, since non-visual data is not available for previously-unseen video footages.

must-links: the two linked samples are in the same cluster; and cannot-links: the two linked samples are from two different clusters. In our multi-source data context, these pairwise constraints are generated from all non-visual data sources. Specifically, first, we computed individual similarity matrices from each non-visual source and averaged them for getting the fused pairwise similarity measure between video samples. The top-k highest and lowest pairwise similarity values were then used to generate must-links and cannot-links respectively. We set $k = \beta * n * (n-1) * 10^{-5}$ where n is the number of training samples, whilst β was cross-validated in a range between 1 and 10, (i.e. k lies in $[339, 3390]$ on TISI, $[881, 8810]$ on ERCe), and the best results were utilised for comparison in our evaluation. Once the Mahalanobis distance metric was learned from both the visual feature data and the generated pairwise links using the algorithm proposed in (Xing et al, 2002), COP-Kmeans (Wagstaff et al, 2001) was employed along with pairwise links as well as the learned metric to obtain the final clusters of video data.

6. *VNV-MSC-Forest-hard*: a variant of our model using hard cluster assignment strategy for inferring semantic tags of previously-unseen samples (Section 4.2), to highlight the effectiveness of the proposed tree structure based tag inference algorithm.
7. *VT-MSC-Forest*: a variant of our model using only temporal information and visual data. In order to show the exact effectiveness of exploiting non-visual data, the weight ratio between visual data and time retains the same as in VNV-MSC-Forest with the only difference of discarding non-visual data during model training.
8. *VPNVρ-MSC-Forest*: a variant of our model but with ρ% of training samples having arbitrary number of missing non-visual types, to evaluate the robustness of MSC-Forest in coping with partial/missing non-visual data.

Implementation details - The clustering forest size γ_{clust} was set to 1000, including both the conventional forest and the proposed MSC-Forest. We observed a slight increase in performance given a larger forest size, which agrees with (Criminisi and Shotton, 2012). The training set X_t of the tth MSC-tree was obtained by performing random selection with replacement from the augmented data space (Figure 3(b)). We set $m_{\text{try}} = \sqrt{d}$ with d the data feature dimension (Eqn. (2)). This is typically practised (Breiman, 2001). We employed linear data separation (Criminisi and Shotton, 2012) as the test function for node splitting. We set the same number of clusters across all methods. This cluster number was discovered automatically using the method presented in (Zelnik-manor and Perona, 2004). For each dataset, $\sim 75\%$ out of the total data was utilised for model training, and the remaining was reserved for testing. Additional previously-unseen video data was collected from the Time Square Intersection scene on a separate day for video summarisation.

Table 2: Compare cluster purity in mean entropy. Lower is better.

Dataset	TISI		ERCe
$p(\mathbf{y}\|c)$	traffic speed	weather	event
VO-Forest (Breiman, 2001)	0.8675	1.0676	0.0616
VNV-Kmeans (Jain, 2010)	0.9197	1.4994	1.2519
VNV-Forest (Breiman, 2001)	0.8611	1.0889	0.0811
VNV-AASC (Huang et al, 2012)	0.7217	0.7039	0.0691
VNV-COP-Mahal (Xing et al, 2002)	0.8523	1.2301	1.0685
VT-MSC-Forest	0.7275	0.9577	0.0580
VNV-MSC-Forest	0.7262	**0.6071**	**0.0024**
VPNV10-MSC-Forest	**0.7190**	0.6261	**0.0024**
VPNV20-MSC-Forest	0.7283	0.6497	0.0090

6 Evaluations

6.1 Multi-Source Clustering

To evaluate the effectiveness of different clustering models for multi-source video clustering, we compared the quality of their clusters formed on the training dataset. For determining clustering quality, we quantitatively measured the mean entropy (Zhao and Karypis, 2004) of non-visual distributions $p(y_i|c)$ (Eqn. (10)) associated with training data clusters to evaluate how coherent video content are partitioned, assuming all methods have access to non-visual data during the entropy computation.

It is evident from Table 2 that our VNV-MSC-Forest achieves the best cluster purity on both datasets[8]. Despite that there are gradual degradations in clustering quality when we increase the non-visual data missing proportion, overall the VNV-MSC-Forest model copes well with partial/missing non-visual data. With no aid of non-visual tag information, VT-MSC-Forest forms much worse clusters. Whilst the superiority of VT-MSC-Forest over VO-Forest suggests the effectiveness of temporal information with MSC-Forest. Inferior performance of VO-Forest to VNV-MSC-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 Kmeans, AASC, and COP-Mahal perform much poorer in comparison to MSC-Forest. The results suggest the proposed joint information gain criterion (Eqn. (4)) is more effective in handling heterogeneous data than the conventional clustering models.

For qualitative comparison, we show examples in Figure 6 using the TISI dataset for detecting 'sunny' weather. It is evident that only VNV-MSC-Forest is able to provide coherent video grouping, with only slight decrease in clustering purity given partial/missing non-visual data. Other

[8] VNV-MSC-Forest-hard shares the same clusters as VNV-MSC-Forest.

Fig. 6: Qualitative comparison on cluster quality on TISI. A key frame of each video is shown. (X/Y) in brackets: X = the number of clips with sunny weather; Y = the total number of clips in a cluster. The frames inside the red boxes are inconsistent clips in a cluster.

methods including VNV-AASC result in a large cluster either leaving out some relevant clips or including many non-relevant ones, with most of them under the influence of strong artificial lighting sources. These non-relevant clips are visually 'close' to sunny weather, but semantically not. The VNV-MSC-Forest model avoids this mistake by correlating both visual and non-visual sources in an information theoretic sense.

6.2 Video Tagging

Generating video summary with semantic interpretations requires accurate tag prediction. In this experiment we compared the performance of different methods in inferring semantic tags given previously-unseen clips extracted from long videos. The proposed tagging algorithm (Section 4.2) is used for VO-Forest, VT-MSC-Forest, VNV-MSC-Forest, and VPNV10/20-MSC-Forest, whilst nearest neighbour (NN) strategy for the others. For quantitative evaluation, we manually annotated 3 weather conditions (sunny, cloudy and rainy) and 4 traffic speeds on TISI previously-unseen clips, whilst 8 event categories on ERCe previously-unseen clips.

Table 3: Comparison of tagging accuracy on TISI.

(%)	traffic speed	weather
VO-Forest (Breiman, 2001)	27.62	50.65
VNV-Kmeans (Jain, 2010)	37.80	43.14
VNV-Forest (Breiman, 2001)	34.95	43.81
VNV-AASC (Huang et al, 2012)	36.13	44.37
VNV-COP-Mahal (Xing et al, 2002)	26.22	40.03
VNV-MSC-Forest-hard	32.86	49.59
VT-MSC-Forest	35.99	54.47
VNV-MSC-Forest	35.77	**61.05**
VPNV10-MSC-Forest	37.99	55.99
VPNV20-MSC-Forest	**38.05**	54.97

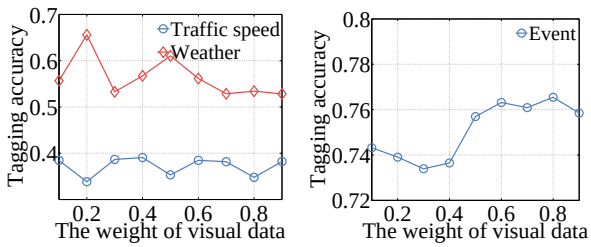
Tagging video by weather and traffic conditions - The experiment was conducted on the TISI outdoor dataset. It

is observed that the performance of different methods (Table 3) is largely in line with their performance in data clustering (Section 6.1). Poor result of tagging traffic conditions is yielded by VO-Forest. This suggests the significance of exploiting non-visual data during model training. It is also seen from Figure 7 that VNV-MSC-Forest not only outperforms other baselines in isolating the sunny weather, but also performs well in distinguishing visually ambiguous cloudy and rainy weathers. In contrast, both VNV-Kmeans and VNV-AASC mistake most of the 'rainy' scenes as either 'sunny' or 'cloudy', as they can be visually similar. Interestingly, the poorest tagging results are obtained by VNV-COP-Mahal where non-visual data is alternatively used as side information for generating pairwise constraints over video samples. The potential reasons include (1) COP-Mahal assumes completely-accurate pairwise links, which however is largely invalid in our context due to the intrinsic noisy nature of non-visual data sources; (2) the errors in pairwise constraints can be propagated during the clustering process of COP-Kmeans and therefore is likely to further worsen the cluster solution and finally the tagging accuracy. This reflects the significant difficulty of jointly learning inherently heterogeneous and inaccurate visual and non-visual data as aforementioned, and in turn the advantages of the proposed joint information gain formulation over existing competitive algorithms.

Tagging video by activity events - Tagging semantic events was tested using the ERCe dataset. By VO-Forest, poor results (Table 4 and Figure 8) are obtained especially on 'Accommodation Service', which involves only subtle activity patterns, i.e. students visiting particular rooms, suggesting using visual data alone is not sufficient to detect such visually subtle events. VT-MSC-Forest over-fits to 'Cleaning' event, therefore performs poorly on 'Student Orientation' event.

Due to the typical high-dimension of visual sources compared to non-visual data, the latter is often overwhelmed by the former in representation. VNV-Kmeans severely suffers from this problem as its most predictions are biased to 'No Scheduled Event' that is more common and frequent visually. This suggests that this distance-based clustering is poor in handling the heteroscedasticity and dimension discrepancy problems in learning heterogeneous data. VNV-AASC attempts to circumvent these problems by seeking for an optimal combination of affinity matrices derived independently from distinct data sources. However this is proved challenging, particularly when each source is inherently noisy and inaccurate. As an alternative way of utilising non-visual data, VNV-COP-Mahal yields again the lowest overall accuracy. This further shows the unsuitability of COP-Mahal in learning ambiguous heterogeneous data due to its stringent assumption on the availability of accurate and reliable pairwise links and the lack of noisy data handling mechanism. In contrast, the proposed MSC-Forest correlates different sources via a joint information gain criterion to effectively alleviate these problems, leading to more robust and accurate tagging performance. Again, VPNV10/20-MSC-Forest perform comparably to VNV-MSC-Forest, further validating the robustness of MSC-Forest in tackling partial/missing non-visual data with the proposed adaptive weighting mechanism (Section 3.2.1).

Interestingly, in some cases, VPNV10/20-MSC-Forest models even outperform VNV-MSC-Forest slightly. We observe that this can be caused by missing noisy non-visual data, which may lead to better results. Overall, the performance difference is marginal and the results demonstrate that MSC-Forest provides stable tagging results across both datasets.



(a) TISI. (b) ERCe.

Fig. 9: The average tagging accuracy against varying visual data weight α_v in Eqn. (4).

Evaluating α sensitivity - We analyse the relative significance of visual data against non-visual and temporal data by varying its weight α_v (Eqn. (4)) in MSC-Forest during model training. The average tagging accuracy is utilised as performance measure criterion. It is observed from Figure 9 that setting $\alpha_v = 0.5$ achieves satisfactory results for both datasets. This observation suggests that visual and non-visual data are almost equally informative. This setting of α is adopted throughout our experiments.

6.3 Semantic Video Summarisation

In this experiment, we follow the method described in Section 4, and show that the learned model MSC-Forest can be easily extended to produce compact yet meaningful video summary of previously-unseen video footage, e.g. from the Time Square Intersection scene, with automatically generated semantic tags. Despite captured from the same scene as the TISI dataset, this previously-unseen video is challenging in that it contains a number of events not seen before (e.g. scaffolding event), with very different weather and traffic conditions. It is interesting to examine how well the multi-

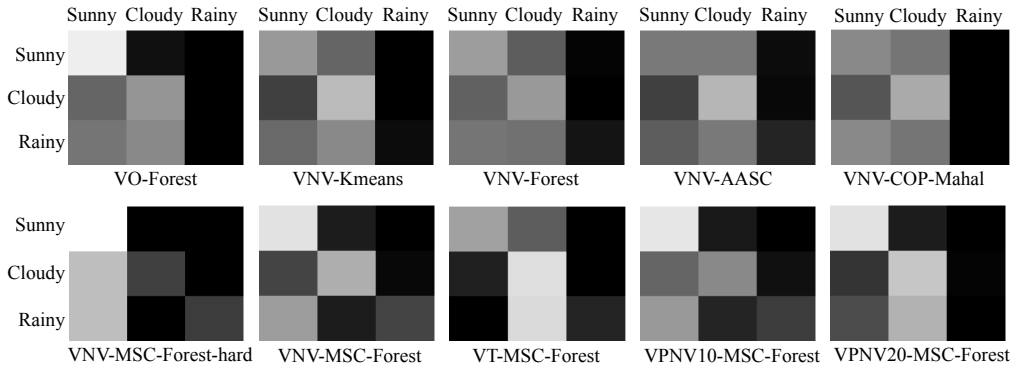


Fig. 7: Weather tagging confusion matrices (TISI dataset).

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

(%)	NoEvt	Cln	CrF	GunFrm	GrStd	SchlCpt	AccSvc	StdOrt	*Average*
VO-Forest (Breiman, 2001)	79.48	39.50	94.41	74.82	92.97	82.74	00.00	60.94	65.61
VNV-Kmeans (Jain, 2010)	**87.91**	19.33	59.38	44.30	46.25	16.71	00.00	09.77	35.45
VNV-Forest (Breiman, 2001)	32.47	30.25	65.46	45.77	41.25	33.15	13.70	33.59	36.96
VNV-AASC (Huang et al, 2012)	48.51	45.80	79.77	84.93	96.88	89.40	21.15	38.87	63.16
VNV-COP-Mahal (Xing et al, 2002)	41.98	**71.43**	54.61	15.07	21.88	00.00	00.24	00.00	25.65
VNV-MSC-Forest-hard	81.25	41.60	70.07	60.48	84.22	82.88	10.82	47.85	59.89
VT-MSC-Forest	57.43	70.17	91.45	79.96	**99.22**	90.08	00.00	43.75	66.50
VNV-MSC-Forest	55.98	41.28	**100.0**	83.82	97.66	99.46	**37.26**	88.09	75.69
VPNV10-MSC-Forest	47.96	46.64	**100.0**	**85.29**	97.66	**99.73**	**37.26**	**92.38**	75.87
VPNV20-MSC-Forest	55.57	46.22	**100.0**	**85.29**	95.78	99.59	37.02	88.09	**75.95**

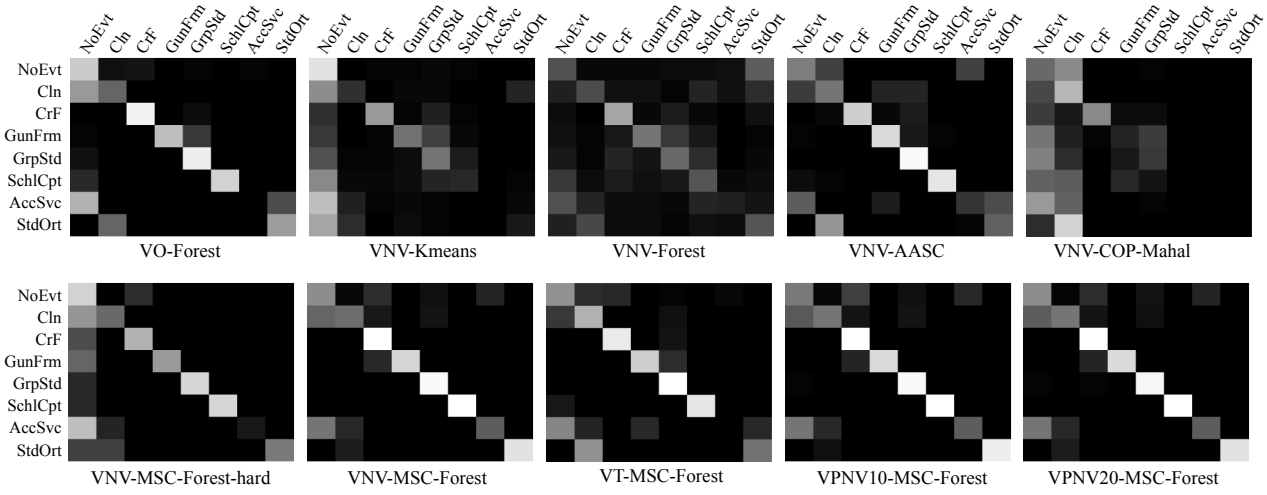


Fig. 8: Event tagging confusion matrices (ERCe dataset).

source model could generalise for drawing meaningful summarisation given such unexpected disparities.

6.3.1 A Quantitative Evaluation on Summary Quality

Measuring the quality of video summary quantitatively is non-trivial since there is no formal definition in the literature. In this study, we employ a *coverage* metric – an ideal summary should cover as many events of interest as possible[9]. More precisely, given a video summary $\mathcal{V}$, its coverage is defined as $\widetilde{c} = \frac{n_{\text{covered}}}{n_{\text{all}}} \left(\frac{\max_i |\mathcal{V}_i|}{|\mathcal{V}|} \right)$, where n_{covered} and

[9] The event of interest is analogous to important objects/regions in (Lee et al, 2012).

n_{all} represent the number of covered and all events of interest, respectively. The $|\mathcal{V}|$ is the length of the current summary, whilst $\max_i |\mathcal{V}_i|$ represents the maximum length of all comparative synopses. The term $\left(\frac{\max_i |\mathcal{V}_i|}{|\mathcal{V}|}\right)$ thus penalises a summary with longer length. Higher coverage is better, implying lower redundancy.

In order to generate unbiased ground truth of event of interest, we asked 10 annotators to watch the previously-unseen video carefully and label each video clip with arbitrary event tags. Although these event tags were produced independently in a somewhat subjective manner, the repetition of similar tagging among different annotators is high, e.g. most annotators labelled 'unloading scaffolding tubes', 'policemen on-duty', as events of their interest. Thus, we formed the ground truth with events that were agreed by over 50% of the annotators. The final ground truth consists of 12 events (Figure 10).

Given the ground truth, we compared the quality of summary generated using the proposed multi-source MSC-Forest with the following baseline methods: (1) Uniform-Sampling: a straightforward way of summarising video by uniformly sampling video clips over time, assuming key events are distributed evenly (Truong and Venkatesh, 2007; Lee et al, 2012). (2) Sufficient-Change: a type of classical summarisation strategy generic to video category (Zhang et al, 1997; Kim and Hwang, 2002; Truong and Venkatesh, 2007). The idea is to select the clip sufficiently different from the previous key clip, e.g. using a pre-defined threshold to decide the change sufficiency. Therefore, the extracted key clips may provide a more diverse and complete summary of the source video. The threshold can be estimated based on the number of key clips. For the distance metric, we adopt L1-norm and L2-norm to measure pairwise similarity between clips in our experiment. (3) VO-Forest: the conventional random forest (Breiman, 2001) that exploits visual features alone. (4) LiveLight (Zhao and Xing, 2014): a dictionary learning based method that considers video summarisation as a sparse coding problem. This aims to encourage the generated summary video to cover sufficiently diverse content with less redundancy. In this sense, LiveLight shares a similar principle to that of "Sufficient-Change" models (both L1 and L2 in Table 5) but with a more sophisticated summarisation algorithm.

More specifically, for VO-Forest and MSC-Forest, we applied the summarisation pipeline described in Section 4 for summary composition. As the code for LiveLight is not publically accessible, we implemented this model by using the SPAMS solver (Mairal et al, 2010) based on the details provided in (Zhao and Xing, 2014). In particular, we fixed the dictionary size to 200 and learned the initial dictionary with the beginning 10 video clips. We cross-validated the threshold of reconstruction error in the range from 0 to 1 for on-line dictionary update, and the best result was utilised for comparison. The summary video was composed using video clips with the highest reconstruction errors. For the remaining methods, we generated the respective video summary via setting a duration similar to the summary by MSC-Forest. Note that non-visual information are not available during the summarisation stage. Hence, for clustering based models, the quality of a summary essentially ties to the purity and coherency of video clusters discovered using different methods.

The results are shown in Figure 10 and Table 5. It is evident that the MSC-Forest model achieves higher event coverage than the baselines. This is in large due to the MSC-Forest's ability for latent data structure discovery (Section 6.1). To reveal concrete reasons on the summarising performance difference, for the same previously-unseen samples $\mathbf{x}^*$ with event of interest, e.g. parcel delivery, we compared the assigned clusters: c^*_{vnv} by our model and c^*_{vo} by VO-Forest. It is found that samples in c^*_{vnv} are visually consistent each other and the majority share some similarity with $\mathbf{x}^*$, e.g. someone standing at the edge of pathway; whilst cluster c^*_{vo} is much larger with no obvious visual commonality over its cluster members. Uniform-Sampling performs poorly since the assumption of uniform event distribution is often invalid. Sufficient-Change is inferior to our model since the visual data distance/similarity measure can be inaccurate and less meaningful due to the challenging semantic gap problem. Owing to the basis component learning strategy and in turn more advanced visual change detection, the LiveLight model can locate more accurately events-of-interest than the non-learning based Sufficient-Change methods. However, the LiveLight model is still inferior to the proposed VNV-MSC-Forest method (Table 5). The plausible reasons are twofold: (1) visual observation obtained from crowded public spaces is often ambiguous and noisy, which makes the learned dictionary unreliable/inaccurate and thus ineffective to model such dynamic visual patterns; (2) to bridge low-level visual features and high-level interpretation is a long-standing challenge and visual-data-only based modelling is typically insufficient to overcome this semantic gap. By jointly learning and correlating both visual and non-visual sources, our VNV-MSC-Forest model shows its superiority and advantage in mitigating this difficulty.

6.3.2 A User Study on Summary Quality

We conducted a user study to examine if the non-visual tags inferred using the MSC-Forest model could complement the unilateral perspective offered by pure visual summary alone. We showed two video summaries to 10 volunteers: (i) a pure visual summary, and (ii) the same summary but enriched with semantic tags inferred using the proposed multi-source

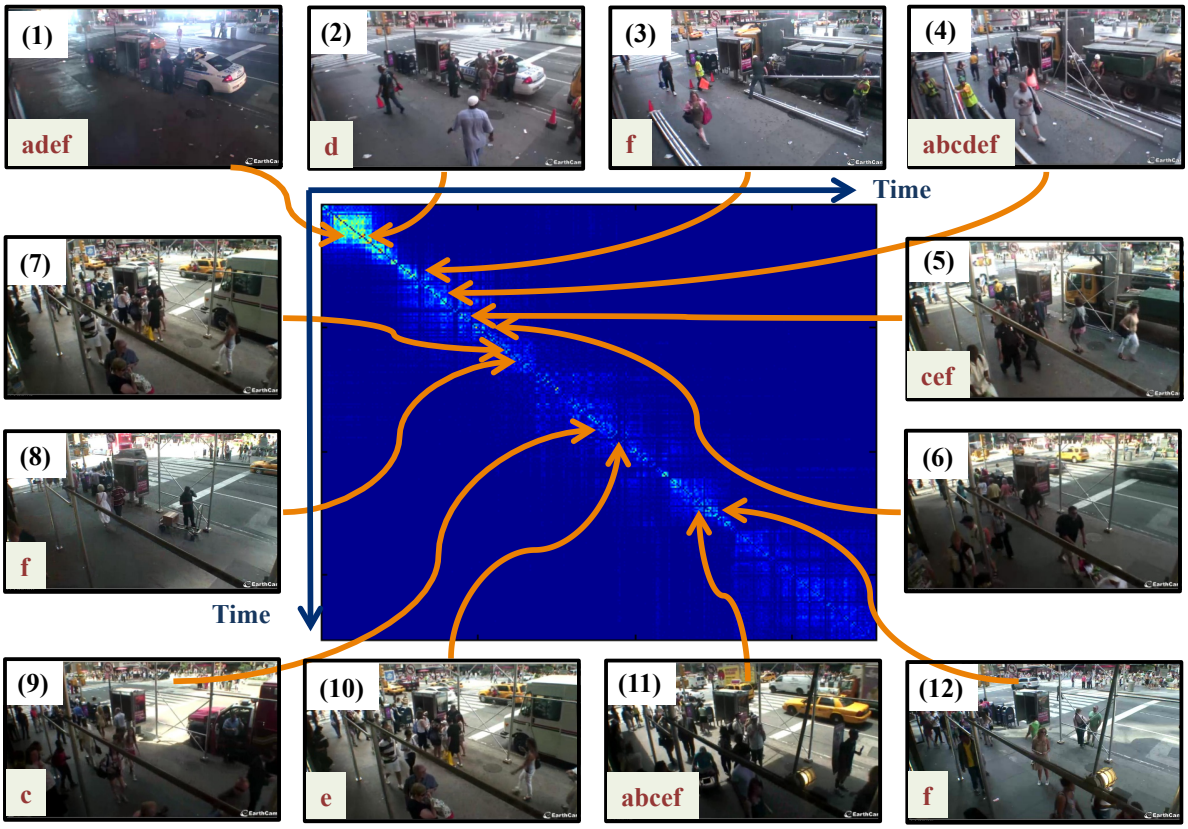


Fig. 10: The multi-source affinity matrix constructed by our model, along with key frames corresponding to ground truth events of interest: (1) policemen on-duty, (2) blocking pathway, (3) workers unloading scaffolding tubes, (4)-(6) different stages of scaffolding, (7)(9)(10) van parking aside, (8) parcel delivery, (11)(12) loitering events. The event covered by some particular method is indicated on the left-bottom corner of key frame with their ID defined as: (a) Uniform-Sampling; (b) Sufficient-Change (L1); (c) Sufficient-Change (L2); (d) VO-Forest; (e) LiveLight; (f) VNV-MSC-Forest.

Table 5: Quantitative comparison of summary. Length = video clip number in summary. Event No. = the number of event-of-interest included in summary.

Method	Length	Event No.	Coverage
Uniform-Sampling	28	3	25.9%
Sufficient-Change(L1)	29	2	16.7%
Sufficient-Change(L2)	29	4	33.3%
VO-Forest	21	3	34.5%
LiveLight	28	5	40.2%
VNV-MSC-Forest (**Ours**)	28	**7**	**60.4%**

model[10]. The tagged summary is shown in Figure 11. Each volunteer was asked to compare and rate the two summaries based on their preference. It is worth pointing out that passing the user test is challenging because providing additional non-visual tags to summary is not necessarily better than none. Tags that correlate poorly with visual context could even jeopardise user experience.

It is evident from Figure 12 that visual summary augmented with non-visual tags was well accepted by all participants over the conventional visual-only summary. A follow-up survey with the volunteers reveals several interesting reasons of their selection. Many volunteers found that the inferred non-visual tags were valuable in providing auxiliary context to achieve better global situational awareness. In particular, the tags helped them to 'connect the dots' and making sense of the previously-unseen (and likely unfamiliar) video footages. Some other volunteers credited the additional non-visual tags in focusing their attention on particular events, and helping them in spotting 'outliers' of interest.

This user study provides an independent means to analyse and validate the usefulness of visual summarisation with auto-tag inference of previously-unseen video footages without a priori semantics or meta-data, mostly typical of surveillance videos. It also shows the effectiveness of the proposed model for mapping multi-source non-visual information to

[10] The inferred non-visual tags include weather, traffic conditions, and typicality. The typicality tag, i.e. *usual* and *interesting*, of each clip, is computed based on the size of their assigned clusters (Figure 4(c)). Clips assigned to the top 20% smallest clusters are treated as 'interesting'.

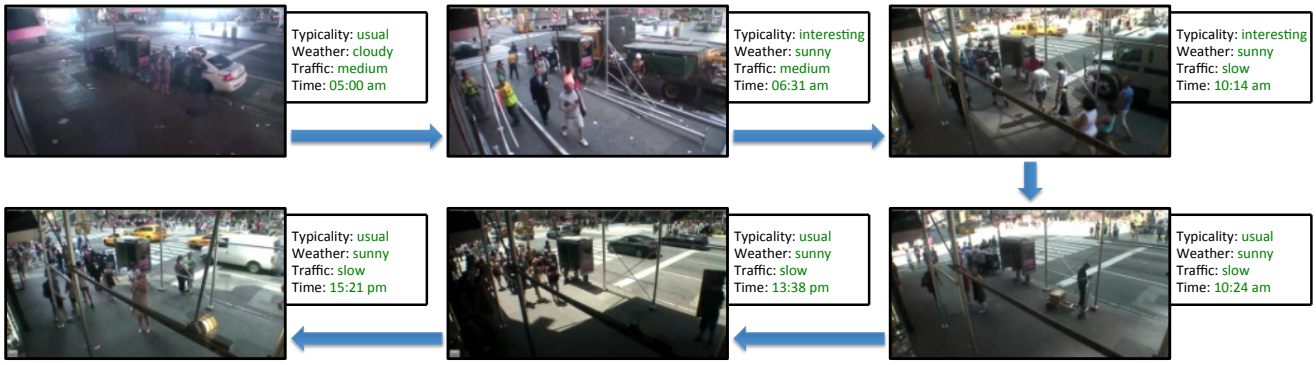


Fig. 11: A storyboard version of our video summary enriched with non-visual tags.

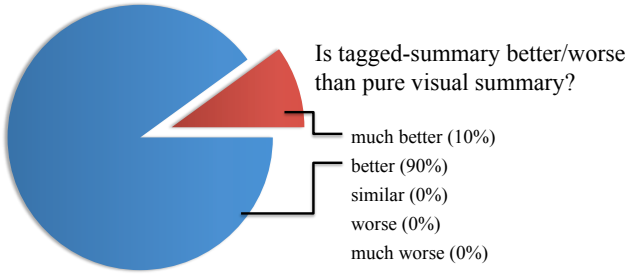


Fig. 12: User study: tagged *versus* pure-visual summary.

unstructured and previously-unseen video data in automatic tagging and summarisation of the videos.

6.4 Multi-Source Model Visualisation

The superior performance of VNV-MSC-Forest can be better explained by examining more closely the capacity of MSC-Forest in uncovering and exploiting the intrinsic correlation among different visual sources and more critically among visual and non-visual sources. This indirect correlation among heterogeneous sources results in well-structured decision trees, subsequently leading to more consistent data clusters and more accurate semantics inference. The details of computing the multi-source correlation are presented in Appendix A. Here we show an example multi-source correction revealed by our MSC-Forest for model visualisation purpose.

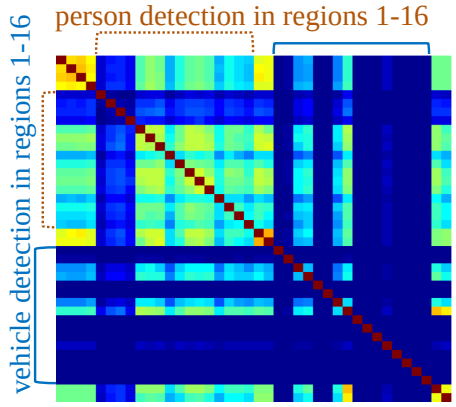
Intuitively, vehicle and person counts should correlate in a busy scene like TISI. Our MSC-Forest discovered this correlation (see Figure 13(a)), so the less reliable vehicle detection from distance against a cluttered background, could enjoy a latent support from more reliable person detection in regions 5-16 close to the camera view.

Moreover, visual sources also benefit from correlated support from non-visual data through our cross-sources information gain optimisation (Eqn. (4)). An example is the intuitive correlation between traffic speed and visual appearance, e.g. slow traffic speed often corresponds to crowded scenarios with a large quantity of pedestrians and vehicles whilst fast traffic speed to sparse people and cars. Such cross-source correlation can be captured by our MSC-Forest, as observed in Figure 10(b) that the vehicle detection responses over road area present a stronger interaction with traffic speed data than those on walk path where vehicles should not appear. In other words, vehicle detection features of road area are preferred over those on walk path in node splitting due to larger induced joint information gain (Eqn. (4)), which is clearly desired. This discovered correlation is further exploited by MSC-Forest during the node splitting optimisation process and thus facilitates the separation of different crowdedness levels of visual data. This leads to better clusters and eventually benefits video summarisation.

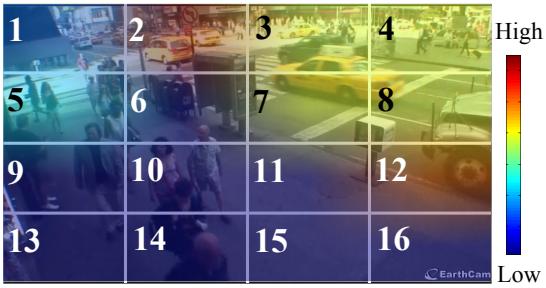
6.5 Computational Costs and Model Complexity

We examined the computational costs for training the proposed MSC-Forest, in comparison to the conventional forests. Time is measured on a Windows PC machine with a dual-core CPU @ 2.66 GHz, 4.0GB RAM, with C++ implementation. Only one core is utilised for training each forest. We recorded the model training time under the same experimental setting as stated in Section 5. It is observed from Table 6 that the training cost of a MSC-Forest model is significantly lower than that of learning conventional forests. In particular, VNV-MSC-Forest records a reduced training time by 14.4% and 17.1% on TISI, and 64.1% and 64.4% on ERCe, when compared with VO-Forest and VNV-Forest, respectively. We observed similar trend on the model inference time.

The lower computational cost of MSC-Forest is owing to its shallow and balanced trees, thanks to the additional non-visual and temporal information during tree optimisation. To make this concrete, we showed in Table 6 the averaged tree fan-in $\varpi^* = \frac{1}{\gamma_{\text{clust}}} \sum_t^{\gamma_{\text{clust}}} \varpi(t)$ of different forest models. A forest with shallow and balanced trees tend to have a small ϖ^* (see Section 3.2.2 for a discussion on tree fan-in). In addition, we also profiled the length of path (from root



(a) Correlations between visual data.



(b) Correlation between vehicle detection and traffic speed.

Fig. 13: The discovered multi-source correlations by our MSC-Forest on TISI.

to leaf node) traversed by training samples. A shallow and balanced tree tends to have shorter path length. The distributions depicted in Figure 14 suggest that MSC-Forest has a shallower and more balanced tree topology than that of conventional forests. It is worth pointing out that despite the shallower structure, MSC-Forest outperforms other models in our clustering and tagging experiments.

Table 6: Random forest model training complexity. Lower is better. TT = Training Time (unit is second).

Dataset	TISI		ERCe	
-	TT	ϖ^*	TT	ϖ^*
VO-Forest	10306	109392	21831	359247
VNV-Forest	10646	108865	22015	359364
VNV-MSC-Forest	**8823**	**91316**	**7845**	**137620**

7 Conclusion and Future Work

We have presented a novel unsupervised multi-source learning model for video summarisation. Specifically, we introduced a joint information gain function for discovering and exploiting latent correlations among independent heterogeneous data sources. The function naturally copes with diverse types of data with different representations, distributions, and dimensions. Importantly, our model is capable of tolerating partial and missing non-visual data, lending it well for automatic semantic tag inference on previously-unseen video footages and for video summarisation. Furthermore, the proposed joint optimisation encourages more compact decision trees, leading to more efficient model training and semantic tag inference. Extensive comparative experiments have demonstrated the advantages of the proposed multi-source video clustering model over existing visual-only models, for both discovering latent video clusters and inferring non-visual semantic tags on previously-unseen video footages. A comprehensive user study was carried out to validate independently the effectiveness of deploying the proposed model for generating contextually-rich and semantically-meaningful video summary.

The proposed model is not limited to surveillance-type videos but can be generalised to other types of unstructured and un-tagged consumer videos or egocentric videos, if 3D camera motion-invariant features or egocentric features (Lee et al, 2012) are adopted. For future work, we will consider generalising/transferring a learned model to new scenes that are significantly different from the training environments. This can be partly addressed by utilising intermediate data representations such as attributes.

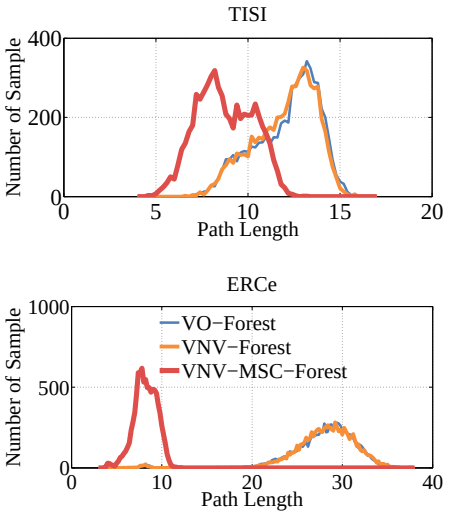


Fig. 14: Comparing tree path length statistics. The same legend is used for both charts.

A Quantifying Correlation between Sources

Quantifying latent correlation between different sources gives insights into their interactions in forming coherent video groupings. This can be done once a MSC-Forest is trained. To quantify between-source correlation, we first estimate correlation among their constituent features.

Visual-visual feature correlation - Visual-visual feature correlation is typically quantified based on their similarity in inducing split node partitions L and R (Breiman, 2001). In particular, given a split node s and its final optimal split, say L_ν and R_ν by feature ν. ¿From Eqn. (2), we recall that this feature ν is selected out from the m_{try} randomly sampled features $F^s = \{f_1, \ldots, f_{m_{\text{try}}}\}$. Let $\tau \in F^s \setminus \nu$ and its optimal left-right partitions be L_τ and R_τ respectively. The node-level correlation between features ν and τ is then defined as

$$\lambda_f(\nu,\tau) = \frac{p_\nu - (1 - \frac{|L_\nu \cap L_\tau|}{|L_\nu \cup R_\nu|} - \frac{|R_\nu \cap R_\tau|}{|L_\nu \cup R_\nu|})}{p_\nu}, \tag{16}$$

where $p_\nu = \min(\frac{|L_\nu|}{|L_\nu|+|R_\nu|}, \frac{|R_\nu|}{|L_\nu|+|R_\nu|})$, thus $p_\nu \in (0, \frac{1}{2}]$. With Eqn. (16) we assign a strong correlation ($\lambda_f(\nu,\tau) = 1$) to a feature pair (ν, τ) if they produce the same data partition, whilst a weak correlation ($\lambda_f(\nu,\tau) \leq -1$) when their partitions have no overlaps. For simplicity we let $\lambda_f(\nu,\tau) = \max(\lambda_f(\nu,\tau), 0)$ such that $\lambda_f(\nu,\tau)$ lies in the range of $[0, 1]$. The final visual-visual feature correlation $\lambda(\nu,\tau)$ is obtained via

$$\lambda(\nu,\tau) = \frac{1}{\gamma_{\text{clust}}} \sum_{t=1}^{\gamma_{\text{clust}}} \left[\frac{1}{N^t_{(\nu,\tau)}} \sum_{k}^{N^t_{(\nu,\tau)}} \lambda_f(\nu,\tau) \right], \tag{17}$$

where $N^t_{(\nu,\tau)}$ refers to the number of sampling co-occurrences of a feature pair (ν, τ) during the splitting process of a MSC-tree t.

Visual-nonvisual feature correlation - Recall that visual and non-visual data play different roles in our MSC-Forest, e.g. the former as splitting features whereas the later as auxiliary information. This difference makes the above equations not applicable to the computation of visual-nonvisual feature correlation since no data split is associated with non-visual features. Instead, we adopt information gain as the visual-nonvisual feature correlation metric. This metric is appropriate in that it also reflects the intrinsic mutual interaction between visual and non-visual features during joint information gain optimisation (Eqn. (4)). Formally, we quantify the node-level correlation between the optimal splitting visual feature ν and a non-visual feature ω as $\lambda_f(\nu,\omega) = \frac{\Delta\psi_\omega}{\psi_{\omega 0}}$ (the non-visual term of Eqn. (4)). The final visual-nonvisual feature correlation $\lambda(\nu,\omega)$ is computed similarly by Eqn. (17).

Correlation between sources - Given between-feature correlation, the final correlation between any two sources ξ_i and ξ_j can then be estimated through

$$\psi(\xi_i,\xi_j) = \frac{1}{|\xi_i||\xi_j|} \sum_{\nu\in\xi_i, \tau\in\xi_j} \lambda(\nu,\tau). \tag{18}$$

References

Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Complex networks: Structure and dynamics. Physics reports pp 175–308

Bosch A, Zisserman A, Munoz X (2007) Image classification using random forests and ferns. In: IEEE International Conference on Computer Vision

Breiman L (2001) Random forests. Machine Learning

Breiman L (2003) Rf/tools: A class of two-eyed algorithms. In: SIAM Workshop, Statistics Department, UC Berkeley

Breiman L, Friedman J, Stone C, Olshen R (1984) Classification and regression trees. Chapman & Hall/CRC

Cai X, Nie F, Huang H, Kamangar F (2011) Heterogeneous image feature integration via multi-modal spectral clustering. In: IEEE Conference on Computer Vision and Pattern Recognition

Caruana R, Karampatziakis N, Yessenalina A (2008) An empirical evaluation of supervised learning in high dimensions. In: International conference on Machine learning

Chan AB, Vasconcelos N (2008) Modeling, clustering, and segmenting video with mixtures of dynamic textures. IEEE Transactions on Pattern Analysis and Machine Intelligence pp 909–926

Chu WS, Song Y, Jaimes A (2015) Video co-summarization: Video summarization by visual co-occurrence. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 3584–3592

Cong Y, Yuan J, Luo J (2012) Towards scalable summarization of consumer videos via sparse dictionary selection. IEEE Transactions on Multimedia 14(1):66–75

Criminisi A, Shotton J (2012) Decision forests: A unified framework. Foundations and Trends in Computer Graphics and Vision pp 81–227

Duin R, Loog M (2004) Linear dimensionality reduction via a heteroscedastic extension of lda: the chernoff criterion. IEEE Transactions on Pattern Analysis and Machine Intelligence pp 732 –739

Felzenszwalb PF, Girshick RB, McAllester DA, Ramanan D (2010) Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence pp 1627–1645

Feng S, Lei Z, Yi D, Li SZ (2012) Online content-aware video condensation. In: IEEE Conference on Computer Vision and Pattern Recognition

Fu Y, Hospedales T, Xiang T, Gong S (2013) Learning multi-modal latent attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence

Gall J, Yao A, Razavi N, Gool LJV, Lempitsky VS (2011) Hough forests for object detection, tracking, and action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence pp 2188–2202

Gong S, Loy CC, Xiang T (2011) Security and surveillance. In: Visual Analysis of Humans, Springer, pp 455–472

Gong Y (2003) Summarizing audiovisual contents of a video program. EURASIP Journal on Advances in Signal Processing pp 160–169

Gygli M, Van Gool HGL (2015) Video summarization by learning submodular mixtures of objectives. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 3090–3098

Gygli M, Grabner H, Riemenschneider H, Van Gool L (2014) Creating summaries from user videos. In: European Conference on Computer Vision, pp 505–520

Heer J, Chi EH (2001) Identification of web user traffic composition using multi-modal clustering and information scent. In: Proc. of the Workshop on Web Mining, SIAM Conference on Data Mining, pp 51–58

Hospedales TM, Li J, Gong S, Xiang T (2011) Identifying rare and subtle behaviors: a weakly supervised joint topic model. IEEE Transactions on Pattern Analysis and Machine Intelligence pp 2451–2464

Huang HC, Chuang YY, Chen CS (2012) Affinity aggregation for spectral clustering. In: IEEE Conference on Computer Vision and Pattern Recognition

Jain AK (2010) Data clustering: 50 years beyond k-means. Pattern Recognition Letters 31(8):651–666

Kang H, Chen X, Matsushita Y, Tang X (2006) Space-time video montage. In: IEEE Conference on Computer Vision and Pattern Recognition

Karydis I, Nanopoulos A, Gabriel HH, Spiliopoulou M (2009) Tag-aware spectral clustering of music items. In: The International Society for Music Information Retrieval, pp 159–164

Khalidov V, Forbes F, Horaud R (2011) Conjugate mixture models for clustering multimodal data. Neural Computation 23:517–557

Khosla A, Hamid R, Lin CJ, Sundaresan N (2013) Large-scale video summarization using web-image priors. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 2698–2705

Kim C, Hwang JN (2002) Object-based video abstraction for video surveillance systems. IEEE Transactions on Circuits and Systems for Video Technology 12:1128–1138

Kim G, Sigal L, Xing EP (2014) Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 4225–4232

Kratz L, Nishino K (2012) Going with the flow: pedestrian efficiency in crowded scenes. In: European Conference on Computer Vision

Lee YJ, Ghosh J, Grauman K (2012) Discovering important people and objects for egocentric video summarization. In: IEEE Conference on Computer Vision and Pattern Recognition

Li W, Mahadevan V, Vasconcelos N (2013) Anomaly detection and localization in crowded scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence

Lin W, Zhang Y, Lu J, Zhou B, Wang J, Zhou Y (2015) Summarizing surveillance videos with local-patch-learning-based abnormality detection, blob sequence optimization, and type-based synopsis. Neurocomputing 155:84–98

Liu B, Xia Y, Yu PS (2000) Clustering through decision tree construction. In: Conference on Information and Knowledge Management

Loy CC, Xiang T, Gong S (2012) Incremental activity modeling in multiple disjoint cameras. IEEE Transactions on Pattern Analysis and Machine Intelligence pp 1799–1813

Lu Z, Grauman K (2013a) Story-driven summarization for egocentric video. In: IEEE Conference on Computer Vision and Pattern Recognition

Lu Z, Grauman K (2013b) Story-driven summarization for egocentric video. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 2714–2721

Mairal J, Bach F, Ponce J, Sapiro G (2010) Online learning for matrix factorization and sparse coding. The Journal of Machine Learning Research 11:19–60

Martin JK (1997) An exact probability metric for decision tree splitting and stopping. Machine Learning pp 257–291

Money AG, Agius H (2008) Video summarisation: A conceptual framework and survey of the state of the art. Journal of Visual Communication and Image Representation pp 121–143

Moosmann F, Nowak E, Jurie F (2008) Randomized clustering forests for image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence pp 1632–1646

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

Oliva A, Torralba A (2001) Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 42:145–175

Perbet F, Stenger B, Maki A (2009) Random forest clustering and application to video segmentation. In: British Machine Vision Conference

Potapov D, Douze M, Harchaoui Z, Schmid C (2014) Category-specific video summarization. In: European Conference on Computer Vision, pp 540–555

Pritch Y, Rav-Acha A, Gutman A, Peleg S (2007) Webcam synopsis: Peeking around the world. In: The IEEE International Conference on Computer Vision

Pritch Y, Rav-Acha A, Peleg S (2008) Nonchronological video synopsis and indexing. IEEE Transactions on Pattern Analysis and Machine Intelligence pp 1971–1984

Schulter S, Leistner C, Wohlhart P, Roth PM, Bischof H (2013a) Alternating regression forests for object detection and pose estimation. In: IEEE International Conference on Computer Vision

Schulter S, Wohlhart P, Leistner C, Saffari A, Roth PM, Bischof H (2013b) Alternating decision forests. In: IEEE Conference on Computer Vision and Pattern Recognition

Shi T, Horvath S (2006) Unsupervised learning with random forest predictors. Journal of Computational and Graphical Statistics pp 118–138

Shotton J, Fitzgibbon A, Cook M, Sharp T, Finocchio M, Moore R, Kipman A, Blake A (2011) Real-time human pose recognition in parts from single depth images. In: IEEE Conference on Computer Vision and Pattern Recognition

Strehl A, Ghosh J (2003) Cluster ensembles—a knowledge reuse framework for combining multiple partitions. The Journal of Machine Learning Research 3:583–617

Sun M, Farhadi A, Seitz S (2014) Ranking domain-specific highlights by analyzing edited videos. In: European Conference on Computer Vision, pp 787–802

Taskiran C, Pizlo Z, Amir A, Ponceleon D, Delp E (2006) Automated video program summarization using speech transcripts. Multimedia, IEEE Transactions on 8:775 – 791

Toderici G, Aradhye H, Pasca M, Sbaiz L, Yagnik J (2010) Finding meaning on youtube: Tag recommendation and category discovery. In: IEEE Conference on Computer Vision and Pattern Recognition

Topchy A, Jain AK, Punch W (2005) Clustering ensembles: Models of consensus and weak partitions. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(12):1866–1881

Truong BT, Venkatesh S (2007) Video abstraction: A systematic review and classification. ACM Transactions on Multimedia Computing, Communications, and Applications

Wagstaff K, Cardie C, Rogers S, Schrödl S, et al (2001) Constrained k-means clustering with background knowledge. In: International Conference on Machine learning, vol 1, pp 577–584

Wang M, Hong R, Li G, Zha ZJ, Yan S, Chua TS (2012) Event driven web video summarization by tag localization and key-shot identification. IEEE Transactions on Multimedia pp 975–985

Wang X, Ma X, Grimson WEL (2009) Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. IEEE Transactions on Pattern Analysis and Machine Intelligence pp 539–555

Wang Z, Zhao M, Song Y, Kumar S, Li B (2010) Youtubecat: Learning to categorize wild web videos. In: IEEE Conference on Computer Vision and Pattern Recognition

Wolf W (1996) Keyframe selection by motion analysis. In: IEEE International Conference on Acoustics, Speech, and Signal Processing

Wu S, Moore BE, Shah M (2010) Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 2054–2060

Xing EP, Jordan MI, Russell S, Ng AY (2002) Distance metric learning with application to clustering with side-information. In: Advances in Neural Information Processing Systems, pp 505–512

Zelnik-manor L, Perona P (2004) Self-tuning spectral clustering. In: Advances in Neural Information Processing Systems

Zhang DQ, Lin CY, Chang SF, Smith JR (2004) Semantic video clustering across sources using bipartite spectral clustering. In: IEEE International Conference on Multimedia and Expo

Zhang H, Wu J, Zhong D, Smoliar SW (1997) An integrated system for content-based video retrieval and browsing. Patten Recognition

Zhao B, Xing EP (2014) Quasi real-time summarization for consumer videos. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 2513–2520

Zhao Y, Karypis G (2004) Empirical and theoretical comparisons of selected criterion functions for document clustering. Machine Learning pp 311–331

Zhu X, Loy CC, Gong S (2013) Constrained clustering: Effective constraint propagation with imperfect oracles. In: IEEE International Conference on Data Mining, pp 1307–1312

Zhu X, Loy CC, Gong S (2014) Constructing robust affinity graphs for spectral clustering. In: Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition, pp 1450–1457

Zhu X, Loy CC, Gong S (2015) Constrained clustering with imperfect oracles. IEEE Transactions on Neural Networks and Learning Systems PP(99)