Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Video Semantic Clustering with Sparse and Incomplete Tags

Jingya Wang, Xiatian Zhu, Shaogang Gong
School of Electronic Engineering and Computer Science
Queen Mary University of London
London E1 4NS, United Kingdom
{jingya.wang, xiatian.zhu, s.gong}@qmul.ac.uk

Abstract

Clustering tagged videos into semantic groups is important but challenging due to the need for jointly learning correlations between heterogeneous visual and tag data. The task is made more difficult by inherently sparse and incomplete tag labels. In this work, we develop a method for accurately clustering tagged videos based on a novel Hierarchical-Multi-Label Random Forest model capable of correlating structured visual and tag information. Specifically, our model exploits hierarchically structured tags of different abstractness of semantics and multiple tag statistical correlations, thus discovers more accurate semantic correlations among different video data, even with highly sparse/incomplete tags.

Introduction

A critical task in video analysis is to automatically discover the intrinsic semantic groups of large quantities of video data (Poppe 2010; Niebles, Wang, and Fei-Fei 2008). However, semantic video clustering by visual feature analysis alone is inherently limited due to the semantic gap between low-level visual features and high-level semantics, particularly under the "curse" of high dimensionality (Beyer et al. 1999). On the other hand, videos are often attached with additional non-visual data, e.g. typically some textual sketch. Such text information include short tags contributed by either users or content providers, such as videos from the YouTube and TRECVID dataset. Exploiting readily accessible textual tags may be beneficial to video clustering as they possess semantic perspectives unique to visual features.

In general, joint learning of visual and text information, two different heterogeneous data modalities, in a shared representational space is non-trivial because: (1) The heteroscedasticity problem (Duin and Loog 2004), that is, disparate data modalities significantly differ in representation (continuous or categorical) and distribution characteristics with different scale and covariance. In addition, the dimensionality of visual data often exceeds that of tag data by a great extent, e.g. thousands vs. tens. Due to this dimensionality discrepancy problem, a simple concatenation of feature spaces will result in an unbalanced representation favourably inclined towards one dominant modality data, leading to suboptimal results. (2) Visual features can be inaccurate and unreliable, due to the inherently ambiguous and noisy visual observation. It is non-trivial to quantify deterministically which feature dimensions are reliable in different videos. (3) The available text tags of video data are often sparse and incomplete. This causes that the visual (with much richer but also noisier and redundant information) and tag (being often sparsely labelled and incomplete) data are not always directly correlated.

Some progress has been made recently on clustering tagged videos. For example, the method of (Zhou et al. 2013) separates the whole task into two independent stages: Tag model learning and video clustering, so that the heteroscedasticity problem can be well mitigated. (Vahdat, Zhou, and Mori 2014) move one step further by not only jointly modelling the correlations between visual features, tags and clusters, but also handling the tag sparseness/incompleteness issue with a tag flipping strategy. Alternatively, data embedding (Chang et al. 2015) can be another solution. However, all these methods are restricted in the following ways.

(I) Tag structure. In these models, tags are organised and used in a flat structure. Nonetheless, different tags may correspond to varying degrees of concept abstractness, e.g. in a hierarchical structure. Ignoring such inherent tag structure may cause degraded data modelling or knowledge loss, as suggested in studies of tag recommendation (she 2008), image segmentation (Zheng et al. 2014), and object recognition (Deng et al. 2014). Typically, existing methods as above assume an accurate hierarchy structure. However, this is not always available, e.g. tag extracted from some loosely structured data source can only provide a rough hierarchy with potentially inaccurate relations, as the meta-data associated with TRECVID videos. Such noisy hierarchy imposes more challenges but still useful if used properly.

(II) Tag correlations. Moreover, tag statistical correlations are not exploited by these clustering models, partly due to model complexity, i.e. incorporating such information into these models is not straightforward nor easy. Correlations are potentially useful as they can offer informative inter-tag constraints. Common tag correlations include: (1) *co-occurrence* which has been used for image annotation (Griffiths and Ghahramani 2005; Chen et al. 2010), and object classification (Deng et al. 2014), and (2) *mutual-*

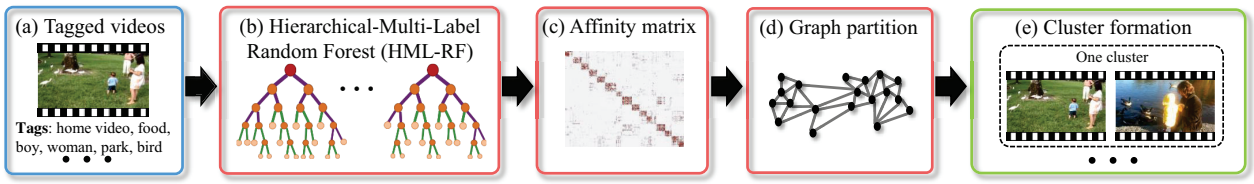

Figure 1: A framework for semantic clustering of tagged videos. (a) Example tagged video; (b) The HML-RF model designed to exploit the inherent tag hierarchy for modelling correlations between ambiguous visual features and sparse tags; (c) HML-RF model based discovery of semantically constrained affinity matrix; (d) Graph-based clustering to discover semantically grouped videos; (e) Resulting video clusters with semantic similarity despite significant visual disparity.

exclusion considered in object detection (Choi et al. 2010; Desai, Ramanan, and Fowlkes 2011), multi-label image annotation (Chen et al. 2011), multi-task learning (Zhou, Jin, and Hoi 2010), object recognition (Deng et al. 2014). Note, our problem differs significantly from these works since: (i) We want to exploit both correlations in an *unsupervised* rather than supervised setting. (ii) Instead of assuming the availability of these tag correlations (Deng et al. 2014), we automatically mine them from sparsely labelled data. Our goal is to exploit automatically extracted unreliable tag correlations for video semantic clustering.

In this work, we develop a model for video semantic clustering by employing both visual features and available sparse/incomplete text tags associated with the videos. We make the following **contributions**: **(I)** We formulate a novel tag-based video clustering method capable of effectively fusing information from ambiguous/noisy visual features and sparse/incomplete textual tags. This is made possible by introducing a new Hierarchical-Multi-Label Random Forest (HML-RF) model with a novel information gain function that allows to model the interactions between visual features and multiple tags simultaneously. Specifically, our model is designed to minimise the uncertainty of tag distributions in an "abstract-to-specific" hierarchical fashion so as to exploit and utilise the intrinsic tag hierarchy structure. **(II)** We introduce a unified tag correlation based algorithm to cope with the tag sparseness/incompleteness problem existing models do not address. Specifically, we formulate a principled way of integrating two automatically mined statistical correlations (co-occurrence and mutual-exclusion) among tags locally during model optimisation. Extensive comparative evaluations on benchmark video/image datasets show the advantages of our model over state-of-the-art methods.

Methodology

Rational for model design. We want to formulate a unified tag-based video clustering model capable of addressing the aforementioned challenges and limitations of existing methods. Specifically, to mitigate the heteroscedasticity and dimension discrepancy problems, we need to isolate different characteristics of visual and tag data, yet can still fully exploit the individual modalities as well as cross-modality interactions in a balanced manner. For handling tag sparseness and incompleteness, we propose to utilise the constraint information derived from inter-tag statistical correlations (Griffiths and Ghahramani 2005; Choi et al. 2010; Deng et al. 2014). To that end, we wish to explore random forest (Breiman 2001; Shi and Horvath 2006; Criminisi 2012) because of: (1) Its flexible training objective function for facilitating multi-modal data modelling and reformulation; (2) The decision tree's hierarchical structures for flexible integration of abstract-to-specific structured tag topology; (3) Its inherent feature selection mechanism for handling data noise. Also, we need to address several shortcomings of the conventional clustering forest (Shi and Horvath 2006; Zhu, Loy, and Gong 2014; 2013a; 2015a), as in its original form it is not best suited for solving our problems in an unsupervised way. Specifically, clustering forest expects a fully concatenated representation as input during model training, it therefore does not allow a balanced utilisation of two modalities simultaneously (the dimension discrepancy problem), nor exploit interactions between visual and tag features. The existing classification forest is also not suitable as it is supervised and aims to learn a prediction function with class labelled training data (usually a single type of tag) (Breiman 2001). Typical video tags do not offer class category labels. However, it is interesting to us that in contrast to the clustering forest, the classification forest offers a more balanced structure for using visual (as split variables) and tag (as semantic evaluation) data that is required for addressing the heteroscedasticity problem by isolating the two heterogeneous modalities during learning.

Approach overview. We want to reformulate classification forest for clustering videos with tags. To that end, we propose a novel *Hierarchical-Multi-Label Random Forest* (HML-RF). Our model goes beyond the classification forest in the following aspects: (1) Employing tags to constrain tree structure learning, rather than learning a generalised prediction function as (Breiman 2001; Criminisi 2012); (2) Introducing a new objective function allowing *acceptance of multi-tags*, *exploitation of abstract-to-specific tag hierarchy* and *accommodation of multiple tag correlations* simultaneously. Instead of learning a classifier, HML-RF is designed to measure pairwise semantic proximity between videos for more accurately revealing the data affinity relationships. These affinity measures over data samples imply their underlying data group/cluster relations that can be then obtained using a standard graph based clustering algorithm. Our tag-based video clustering pipeline is depicted in Fig. 1.

Notations. We consider two data modalities: (1) *Visual modality* - We extract a d-dimensional visual descriptor from the i-th video sample denoted by $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,d}) \in R^d, i = 1, \ldots, n$. All visual features are formed as $X = \{\mathbf{x}_i\}_{i=1}^n$. (2) *Tag modality* - Tags are extracted from the meta-data files associated with videos. We represent m types of binary tag data ($Z = \{1, \ldots, m\}$) attached with the i-th video as $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,m}) \in [0, 1]^m$. All tag data is defined as $Y = \{\mathbf{y}_i\}_{i=1}^n$.

Hierarchical-Multi-Label Random Forest (HML-RF). Let us introduce a HML-RF model, which is an extended hybrid model of classification and clustering forests. The model inputs include visual features $\mathbf{x}$ and tag data $\mathbf{y}$ of video samples (analogous to classification forest), and the output is an estimated affinity matrix $\mathbf{A}$ over input samples X (similar to clustering forest). Specifically, HML-RF contains an ensemble of τ decision trees (HML-trees). Growing a HML-tree t involves a recursive node splitting procedure on randomly sampled data $X_t \subset X$ until some stopping criterion is satisfied. The training of each split node is a process of binary split function optimisation, defined as

$$\mathrm{h}(\mathbf{x}, \mathbf{w}) = \begin{cases} 0 & \text{if } x_{w_1} < w_2, \\ 1 & \text{otherwise.} \end{cases} \quad (1)$$

with two parameters $\mathbf{w} = [w_1, w_2]$: (i) a feature dimension x_{w_1} with $w_1 \in \{1, \ldots, d\}$, and (ii) a feature threshold w_2 The optimal split parameter $\mathbf{w}^*$ is chosen via

$$\mathbf{w}^* = \mathrm{argmax}_{\mathbf{w} \in W} \Delta\psi_{\mathrm{sl}}, \quad (2)$$

where $W = \{\mathbf{w}_i\}_{i=1}^{\nu_{\mathrm{try}}(|S|-1)}$ represents a parameter set over ν_{try} randomly selected features, with S the sample set reaching the node s. The cardinality of a set is given by $|\cdot|$. The information gain $\Delta\psi_{\mathrm{sl}}$ is formulated as

$$\Delta\psi_{\mathrm{sl}} = \psi_s - \frac{|L|}{|S|}\psi_l - \frac{|R|}{|S|}\psi_r, \quad (3)$$

where L and R denote the data set routed into the left l and right r children, and $L \cup R = S$. The uncertainty ψ over the tag distribution can be computed as the Gini impurity (Breiman et al. 1984).

Accommodating multiple tags. Conventional classification forests typically assume single tag (or label) type. In contrast, the new HML-RF can accept multiple types simultaneously by extending the single-tag based information gain function Eqn. (3) to multi-tags for training HML-trees:

$$\Delta\psi_{\mathrm{ml}} = \sum_{i=1}^{m} \Delta\psi_{\mathrm{sl}}^i \quad (4)$$

where $\Delta\psi_{\mathrm{sl}}^i$ is individual information gain computed in the i-th tag by Eqn. (3). Hence, the split functions are optimised in a similar way as supervised classification forests, and semantics from multiple tags are enforced simultaneously. *Discussion:* In the context of tagged video clustering, it should be noted that our way of exploiting tags is not supervised since the tags are not target classes. We call this *structurally-constrained clustering*. Additionally, the interactions between visual features (on which split functions are defined) and tags (used to optimise split functions) are also modelled during learning via identifying the most discriminative features w.r.t. a collection of tags. Importantly, this separation of visual and tag data naturally solves the dimensionality discrepancy challenge. Moreover, HML-RF benefits from the feature selection mechanism inherent to random forest for coping with the noisy visual data problem by selecting the most discriminative localised split functions (Eqn. (1)) over multiple tags simultaneously.

Incorporating tag hierarchy. Eqn. (4) implies that all the tags have similar abstractness in semantics, as all of them are used in every split node (i.e. a flatten structure of tags). To further exploit available structural information in tag hierarchy, we introduce an adaptive hierarchical multi-label information gain function as:

$$\Delta\psi_{\mathrm{hml}} = \sum_{k=1}^{\mu} \left(\prod_{j=1}^{k-1} (1 - \alpha_j) \alpha_k \sum_{i \in Z_k} \Delta\psi_{\mathrm{sl}}^i \right) \quad (5)$$

where Z_k denotes the tag index set of the k-th layer in the tag hierarchy (totally μ layers), with $\cup_{k=1}^{\mu} Z_k = Z$, and $\forall_{j \neq k} Z_j \cap Z_k = \emptyset$. Binary flag $\alpha_k \in \{0, 1\}$ indicates the impurity of the k-th tag layer, $k \in \{1, \ldots, \mu\}$, i.e. $\alpha_k = 0$ when tag values are identical/pure across all the training samples S of split node s in any tag $i \in Z_k$, $\alpha_k = 1$ otherwise. The target layer is k in case that $\alpha_k = 1$ and $\forall \alpha_j = 0, 0 < j < k$.

Discussion: This layer-wise design allows the data partition optimisation to concentrate on the *most abstract* and *impure* tag layer (i.e. the target layer) so that the tag structural knowledge can be naturally embedded into the top-down HML-tree growing procedure in an abstract-to-specific fashion. We shall show the empirical effectiveness of this layer-wise information gain design in our experiments.

Handling tag sparseness and incompleteness. We further improve the HML-RF model by employing tag statistical correlations for addressing tag sparseness problem, as follows: We wish to utilise the dependences among tags to infer missing tags with a confidence measure (continuous soft tags), and exploit them along with labelled (binary hard) tags in localised split node optimisation, e.g. Eqn. (3) and (5).

In particular, two tag correlations are considered: *co-occurrence* - often co-occur in the same video thus positively correlated, and *mutual-exclusion* - rarely simultaneously appear so negatively correlated. They are complementary to each other, since for a particular sample, co-occurrence helps predict the *presence* degree of some missing tag based on another often co-occurrent tag who is labelled, whilst mutual-exclusion can estimate the *absence* degree of a missing tag according to its negative relation with another labelled tag. Therefore, we infer tag positive $\{\hat{y}_{\cdot,i}^+\}$ and negative $\{\hat{y}_{\cdot,i}^-\}$ confidence scores based upon tag co-occurrent and mutual-exclusive correlations, respectively. In our layered optimisation, we restrict the notion of missing tag to videos $S_{\mathrm{miss}} = \{\mathring{\mathbf{x}}\}$ where no tag in the target layer is labelled, and consider cross-layer tag correlations considering that a hierarchy is typically shaped as a pyramid, with more specific tag categories at lower layers where likely

more labelled tags are available. Suppose we compute the correlations between the tag $i \in Z_k$ (the target tag layer) and the tag $j \in \{Z_{k+1}, \ldots, Z_\mu\}$ (subordinate tag layers).
Co-occurrence: We compute the co-occurrence $\varrho_{i,j}$ as

$$\varrho_{i,j} = co_{i,j}/o_j, \tag{6}$$

where $co_{i,j}$ denotes the co-occurrence frequency of tags i and j, that is, occurrences when both tags simultaneously appear in the same video across all samples; and o_j denotes the number of occurrences of tag j over all samples. The denominator o_j here is used to down-weight over-popular tags j: Those often appear across the dataset, and their existence thus gives a weak positive cue of supporting the simultaneous presence of tag i. Once $\varrho_{i,j}$ is obtained, for a potentially missing tag $i \in Z_k$ of $\mathring{\mathbf{x}} \in S_{\text{miss}}$, we estimate its positive score $\hat{y}^+_{\cdot,i}$ via:

$$\hat{y}^+_{\cdot,i} = \sum_{j \in \{Z_{k+1}, \ldots, Z_\mu\}} \varrho_{i,j} y_{\cdot,j} \tag{7}$$

where $y_{\cdot,j}$ refers to the j-th tag value of $\mathring{\mathbf{x}}$. With Eqn. (7), we accumulate the positive support from all labelled subordinate tags to estimate the presence confidence of tag i.
Mutual-exclusion: We calculate this negative correlation as

$$\epsilon_{i,j} = \max(0, r^{-+}_{i,j} - r^-_i)/(1 - r^-_i), \tag{8}$$

where r^-_i refers to the negative sample percentage on tag i across all samples, and $r^{-+}_{i,j}$ the negative sample percentage on tag i over samples with positive tag j. The denominator $(1 - r^-_i)$ is the normalisation factor. Hence, $\epsilon_{i,j}$ is approximated statistically as the relative increase in negative sample percentage on tag i given positive tag j. This definition reflects statistical exclusive degree of tag j against tag i naturally. The cases of $\epsilon < 0$ are not considered since they are already measured in the co-occurrence. Similarly, we predict the negative score $\hat{y}^-_{\cdot,i}$ for $\mathbf{x}$ on tag i with:

$$\hat{y}^-_{\cdot,i} = \sum_{j \in \{Z_{k+1}, \ldots, Z_\mu\}} \epsilon_{i,j} y_{\cdot,j}, \tag{9}$$

Finally, we normalise both $\hat{y}^+_{\cdot,i}$ and $\hat{y}^-_{\cdot,i}$, $i \in Z_p$, into the unit range $[0, 1]$.

Inducing data affinity from a trained HML-RF model. Recall that our HML-RF model is designed to measure pairwise proximity between samples (Fig. 1(c)). This is inspired by clustering forests (Breiman 2001; Shi and Horvath 2006). Specifically, the t-th ($t \in \{1, \ldots, \tau\}$) HML-tree partitions the training samples at its leaves. We assign pairwise similarity 1 for sample pair ($\mathbf{x}_i$, $\mathbf{x}_j$) if they fall into the same leaf, and 0 otherwise. This results in a tree-level affinity matrix $\mathbf{A}^t$. A smooth affinity matrix $\mathbf{A}$ can be obtained through: $\mathbf{A} = \frac{1}{\tau}\sum_{t=1}^{\tau} \mathbf{A}^t$, with τ the tree number of HML-RF. Intuitively, the multi-modality learning strategies of HML-RF enable its data similarity measure to be much more meaningful. This can benefit significantly video clustering using a graph-based clustering method, e.g. spectral clustering (Ng et al. 2002) which is utilised in this work.

Experiments

Dataset. We used the TRECVID MED 2011 dataset (Over et al. 2011) to evaluate the efficacy of the proposed HML-RF model for tag-based video clustering. Particularly, TRECVID MED 2011 includes 2379 video samples from 15 categories: 'board trick', 'feeding animal', etc. We aim to group these videos into the above categories as in (Zhou et al. 2013; Vahdat, Zhou, and Mori 2014). This dataset is challenging for clustering using only visual features, in that videos with the same concepts can present significant variety/dynamics in visual appearance. This necessitates the assistance of other data modalities, e.g. textual meta-data.
Visual features. For TRECVID MED 2011, we used HOG3D features (Klaser, Marszałek, and Schmid 2008) as visual representation of videos. In particular, we first generated a codebook of 1000 words using Kmeans (Jain 2010). With this codebook, we created a histogram feature vector for each video. Finally, the approximated Histogram Intersection Kernel via feature extension (Vedaldi and Zisserman 2012) was adopted to further enhance the expressive capability of visual features.
Tags and hierarchy. Automatically extracted tags from the textual meta-data, e.g. the judgement files associated with video samples, are utilised (Vahdat, Zhou, and Mori 2014). A total of 114 tags were obtained and used in our evaluation. We established the tag hierarchy according to the structure presented in the meta-data files where two levels of abstractness: holistic (e.g. event) and specific (e.g. object) concepts. Averagely, ~ 4 tags were extracted per video, thus very sparse with many unknown missing tags.
Input data modes. For comparison, we tested four modes of input data: (1) ViFeat: Videos are represented by HOG3D visual features; (2) BiTag: Binary tag vectors are used instead of visual features; (3) DetScore (Vahdat, Zhou, and Mori 2014): Tag classifiers (e.g. SVM) are trained for individual tags using the available tags with HOG3D visual features and their detection scores are then used as model input[1]; (4) ViFeat&BiTag: Both the visual and tag data are utilised. More specifically, the two modalities may be combined into one single feature vector (called V&T-cmb), or modelled separately in some balanced way (called V&T-bln), depending on the design nature of specific methods.
Baseline models. we extensively compared our HML-RF model against 10 existing and state-of-the-arts clustering methods: (1) Kmeans (Jain 2010); (2) Spectral Clustering (SpClust) (Ng et al. 2002): For ViFeat&BiTag mode, the averaging over separate normalised affinity matrices of visual and tag data (SpClust-Av) was also evaluated, in addition to the combined single feature (SpClust-Cmb); (3) Affinity Propagation (AffProp) (Frey and Dueck 2007); (4) Clustering Random Forest (ClustRF) (Breiman 2001; Shi and Horvath 2006): It was used to generate the data affinity matrix, followed by SpClust for obtaining the final clusters; (5) Constrained-Clustering Forest (CC-Forest) (Zhu, Loy, and Gong 2013b; 2015b): A state-of-the-art multi-modality data

[1]We only compared the reported results in (Vahdat, Zhou, and Mori 2014) since we cannot reimplement the exact evaluation setting due to the lack of experimental details.

Table 1: Comparing clustering models on TRECVID MED 2011.

Input	Method	Purity	NMI	RI	F1	ARI
Visual	Kmeans(Jain 2010)	0.26	0.19	0.88	0.14	0.08
	SpClust(Ng et al. 2002)	0.25	0.20	0.88	0.15	0.07
	ClustRF(Breiman 2001)	0.23	0.17	0.87	0.14	0.08
	AffProp(Frey and Dueck 2007)	0.23	0.16	0.87	0.14	0.07
	MMC(Xu et al. 2004)	0.25	0.19	0.88	0.14	0.09
BiTag	Kmeans(Jain 2010)	0.51	0.52	0.86	0.30	0.23
	SpClust(Ng et al. 2002)	0.71	0.73	0.93	0.56	0.60
	ClustRF(Breiman 2001)	0.77	0.81	0.94	0.64	0.60
	AffProp(Frey and Dueck 2007)	0.50	0.44	0.87	0.28	0.21
	MMC(Xu et al. 2004)	0.76	0.72	0.95	0.64	0.60
DetScore	Kmeans(Jain 2010)	0.63	0.60	0.93	0.50	-
	SpClust(Ng et al. 2002)	0.82	0.76	0.96	0.69	-
	MMC(Xu et al. 2004)	0.83	0.78	0.96	0.73	-
	L-MMC(Zhou et al. 2013)	0.86	0.82	0.97	0.79	-
V&T-cmb	Kmeans(Jain 2010)	0.51	0.49	0.90	0.34	0.24
	SpClust-cmb(Ng et al. 2002)	0.76	0.74	0.94	0.62	0.66
	ClustRF(Breiman 2001)	0.23	0.17	0.87	0.15	0.08
	AffProp(Frey and Dueck 2007)	0.51	0.46	0.86	0.29	0.21
V&T-bln	SpClust-bln(Ng et al. 2002)	0.75	0.72	0.95	0.62	0.59
	CC-Forest(Zhu, Loy, and Gong 2013b)	0.41	0.33	0.89	0.41	0.19
	AASC(Huang, Chuang, and Chen 2012)	0.30	0.15	0.87	0.13	0.06
	MMC(Xu et al. 2004)	0.79	0.72	0.95	0.66	0.66
	S-MMC(Vahdat, Zhou, and Mori 2014)	0.87	0.84	0.97	0.79	-
	F-MMC(Vahdat, Zhou, and Mori 2014)	0.90	0.88	**0.98**	0.84	-
	HML-RF(Ours)	**0.94**	**0.90**	**0.98**	**0.88**	**0.87**

based clustering forest characterised by joint learning of heterogeneous data; (6) Affinity Aggregation for Spectral Clustering (AASC) (Huang, Chuang, and Chen 2012): A state-of-the-art multi-modal spectral clustering method; (7) Maximum Margin Clustering (MMC) (Xu et al. 2004); (8) Latent Maximum Margin Clustering (L-MMC) (Zhou et al. 2013): An extended MMC model that allows to accommodate latent variables, e.g. tag labels, during maximum cluster margin learning; (9) Structural MMC (S-MMC) (Vahdat, Zhou, and Mori 2014): A variant of MMC model assuming structured tags are labelled on data samples; (10) Flip MMC (F-MMC) (Vahdat, Zhou, and Mori 2014): The state-of-the-art tag based video clustering method capable of handling the missing tag problem, beyond S-MMC.

Evaluation metrics. We adopted five metrics to evaluate the clustering accuracy: (1) *Purity* (Zhou et al. 2013), which calculates the averaged accuracy of the dominating class in each cluster; (2) *Normalised Mutual Information* (NMI) (Vinh, Epps, and Bailey 2009), which considers the mutual dependence between the predicted and ground-truth partitions; (3) *Rand Index* (RI) (Rand 1971), which measures the ratio of agreement between two partitions, i.e. true positives within clusters and true negatives between clusters; (4) *Adjusted Rand Index* (ARI) (Steinley 2004), an adjusted form of RI that additionally considers disagreement; (5) *balanced F1 score* (F1) (Jardine and van Rijsbergen 1971), which uniformly measures both precision and recall. All metrics lie in $[0, 1]$ except ARI in $[-1, 1]$. For each, higher values indicate better performance.

Implementation details. The forest size τ was fixed to 1000 for all forest variants. The depth of each tree was automatically determined by setting the sample number in the leaf node, ϕ, which we set to 3 throughout our experiments. We set $\nu_{\text{try}} = \sqrt{d}$ with d the data feature dimension (Eqn. (2)). For fair comparison, we used the exactly same number of clusters, visual features and tag data in all methods.

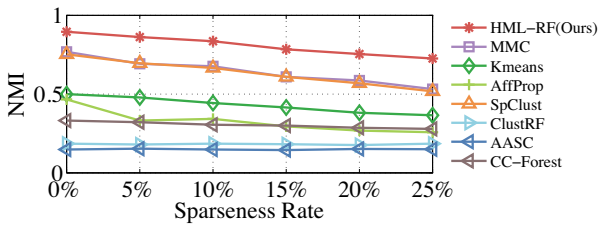

Figure 2: Clustering performance of compared methods at different ratios of tag sparseness. Metric: NMI.

Table 2: Comparing the relative drop of top-3 models, given different tag sparseness rates. Metric: NMI.

Sparseness (%)	10	15	20	25
SpClust(Ng et al. 2002)	0.11	0.19	0.24	0.31
MMC(Xu et al. 2004)	0.12	0.21	0.24	0.31
HML-RF(Ours)	**0.07**	**0.12**	**0.16**	**0.19**

Table 3: Comparing clustering models on NUS-WIDE.

Input	Method	Purity	NMI	RI	F1	ARI
V&T-bln	SpClust-bln(Ng et al. 2002)	0.71	0.73	0.91	0.50	0.46
	CC-Forest(Zhu, Loy, and Gong 2013b)	0.23	0.23	0.86	0.13	0.05
	AASC(Huang, Chuang, and Chen 2012)	0.26	0.10	0.86	0.13	0.05
	MMC(Xu et al. 2004)	0.24	0.11	0.88	0.13	0.06
	HML-RF(Ours)	**0.83**	**0.80**	**0.96**	**0.74**	**0.72**

Clustering Evaluation

We evaluated the effectiveness of different models for tag-based video clustering, using the *full* tag data along with visual features. The results are reported in Table 1. Given the visual features alone, all clustering methods produce poor results, e.g. the best NMI is 0.20, achieved by SpClust. Whereas binary tags provide much more information about the underlying video data structure than visual modality, e.g. all models can double their scores or even more in most metrics. Interestingly, using the detection scores can lead to even better results than the original binary tags. The plausible reason is that missing tags can be partially recovered after using detection scores. When using both modalities, we observe superior results than either single modality with many methods like SpClust, AffProp, MMC. This confirms the overall benefits from jointly learning visual and tag data.

For the performance of individual methods, the proposed HML-RF model evidently provides the best results by a significant margin over the second best Flip MMC in most metrics. This is resulted from the joint exploitation of interactions between visual and tag data, tag hierarchical structure, and tag correlations with a unified HML-RF model, compared to MMC and its variants wherein tags are exploited in a flat organisation and no tag dependences are considered. Kmeans hardly benefits from the combination of visual and tag data, due to its single distance function based grouping mechanism therefore is very restricted in jointly exploiting multi-modality data.

Among all affinity based models, ClustRF is surprisingly dominated by visual data when using visual features & tag as input. This may be because that visual features with large variances may be mistakenly considered as optimum due to larger information gain induced on them. CC-Forest suffers less by separately exploiting the two modalities, but still inferior than HML-RF due to ignoring the intrinsic tag structure and the sparseness challenge. AASC yields much poorer clustering results than HML-RF, suggesting that the construction of individual affinity matrices can lose significant information, e.g. the interactions between the visual and tag data, as well as the tag correlations. AffProp and SpClust also suffer from the heteroscedasticity problem due to the input affinity matrix is constructed from the heterogeneous concatenation of visual and tag data and thus ineffective to exploit the knowledge embedded across modalities and in tag statistical relationships.

Comparing Robustness against Tag Sparseness. We conducted a scalability evaluation against tag sparseness and incompleteness. This is significant since we may have access to merely a small size of tags in many practical settings. To simulate these scenarios, we randomly removed varying ratios ($5 \sim 25\%$) of tag data. We utilised both visual and tag data as model input since most methods can benefit from using both[2]. The most common metric NMI (Jain 2010) was used in this experiment.

The experimental results are compared in Fig. 2. Given less amount of tags, we can observe a clear performance drop trend across most models. Numerically, the relative drops of HML-RF are significantly smaller than those most competitive (top-2) models, e.g. MMC and SpClust (Table 2). Whilst the remaining are slightly affected by tag sparseness, in that these methods are very ineffective in exploiting tag data and thus less sensitive.

Further Evaluation

We additionally evaluated the HML-RF model on the tagged NUS-WIDE image dataset (Chua et al. 2009) with the released visual features and 1000 tags. A subset of 14 randomly selected categories was utilised. We built a 2-layer tag hierarchy by Kmeans. It is evident from Table 3 that HML-RF achieves the best results.

Conclusion

We presented an unsupervised tag based video semantic clustering framework by formulating a novel Hierarchical-Multi-Label Random Forest model for jointly exploiting heterogeneous visual and tag data. The proposed forest model, which is defined by a new information gain function, enables to naturally incorporate tag abstractness hierarchy and effectively exploit multiple tag statistical correlations, beyond modelling the intrinsic interactions between visual and tag modalities. Extensive comparative evaluations on clustering challenging tagged videos and images have demonstrated the advantages of the proposed model over a wide range of existing and state-of-the-art clustering models.

Acknowledgement

We shall thank Arash Vahdat, Guang-Tong Zhou, and Greg Mori for providing us with the visual features and tags for comparative evaluations on the TRECVID MED 2011 dataset (Over et al. 2011).

[2]Structural MMC and Flip MMC models (Vahdat, Zhou, and Mori 2014) were not included in this evaluation due to the difficulties in reconstructing their models from a lack of sufficient implementational details.

References

Beyer, K.; Goldstein, J.; Ramakrishnan, R.; and Shaft, U. 1999. When is nearest neighbor meaningful? In *Database Theory-ICDT99*. Springer. 217–235.

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

Breiman, L. 2001. Random forests. *Machine learning* 45(1):5–32.

Chang, S.; Han, W.; Tang, J.; Qi, G.-J.; Aggarwal, C. C.; and Huang, T. S. 2015. Heterogeneous network embedding via deep architectures. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 119–128.

Chen, X.; Mu, Y.; Yan, S.; and Chua, T.-S. 2010. Efficient large-scale image annotation by probabilistic collaborative multi-label propagation. In *ACM International Conference on Multimedia*, 35–44.

Chen, X.; Yuan, X.-T.; Chen, Q.; Yan, S.; and Chua, T.-S. 2011. Multi-label visual classification with label exclusive context. In *IEEE International Conference on Computer Vision*, 834–841.

Choi, M. J.; Lim, J. J.; Torralba, A.; and Willsky, A. S. 2010. Exploiting hierarchical context on a large database of object categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, 129–136.

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, 48.

Criminisi, A. 2012. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Found. Trends. Comp. Graphics and Vision*.

Deng, J.; Ding, N.; Jia, Y.; Frome, A.; Murphy, K.; Bengio, S.; Li, Y.; Neven, H.; and Adam, H. 2014. Large-scale object classification using label relation graphs. In *European Conference on Computer Vision*. 48–64.

Desai, C.; Ramanan, D.; and Fowlkes, C. C. 2011. Discriminative models for multi-class object layout. *International Journal of Computer Vision* 95(1):1–12.

Duin, R., and Loog, M. 2004. Linear dimensionality reduction via a heteroscedastic extension of lda: the chernoff criterion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(6):732–739.

Frey, B. J., and Dueck, D. 2007. Clustering by passing messages between data points. *Science* 315(5814):972–976.

Griffiths, T., and Ghahramani, Z. 2005. Infinite latent feature models and the indian buffet process. In *Advances in Neural Information Processing Systems*, 475–482.

Huang, H.-C.; Chuang, Y.-Y.; and Chen, C.-S. 2012. Affinity aggregation for spectral clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 773–780.

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

Jardine, N., and van Rijsbergen, C. J. 1971. The use of hierarchic clustering in information retrieval. *Information storage and retrieval* 217–240.

Klaser, A.; Marszałek, M.; and Schmid, C. 2008. A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference*, 275–1.

Ng, A. Y.; Jordan, M. I.; Weiss, Y.; et al. 2002. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, volume 2, 849–856.

Niebles, J. C.; Wang, H.; and Fei-Fei, L. 2008. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision* 79(3).

Over, P.; Awad, G. M.; Fiscus, J.; Antonishek, B.; Michel, M.; Smeaton, A. F.; Kraaij, W.; and Quénot, G. 2011. Trecvid 2010–an overview of the goals, tasks, data, evaluation mechanisms, and metrics.

Poppe, R. 2010. A survey on vision-based human action recognition. *Image and vision computing* 28(6):976–990.

Rand, W. M. 1971. Objective criteria for the evaluation of clustering methods. *American Statistical association* 66(336):846–850.

2008. Personalized recommendation in social tagging systems using hierarchical clustering. In *Proceedings of the 2008 ACM conference on Recommender systems*, 259–266.

Shi, T., and Horvath, S. 2006. Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics* 15(1).

Steinley, D. 2004. Properties of the hubert-arable adjusted rand index. *Psychological methods* 9(3):386.

Vahdat, A.; Zhou, G.-T.; and Mori, G. 2014. Discovering video clusters from visual features and noisy tags. In *European Conference on Computer Vision*. 526–539.

Vedaldi, A., and Zisserman, A. 2012. Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(3):480–492.

Vinh, N. X.; Epps, J.; and Bailey, J. 2009. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *International Conference on Machine learning*, 1073–1080.

Xu, L.; Neufeld, J.; Larson, B.; and Schuurmans, D. 2004. Maximum margin clustering. In *Advances in Neural Information Processing Systems*, 1537–1544.

Zheng, S.; Cheng, M.-M.; Warrell, J.; Sturgess, P.; Vineet, V.; Rother, C.; and Torr, P. H. 2014. Dense semantic image segmentation with objects and attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3214–3221.

Zhou, G.-T.; Lan, T.; Vahdat, A.; and Mori, G. 2013. Latent maximum margin clustering. In *Advances in Neural Information Processing Systems*, 28–36.

Zhou, Y.; Jin, R.; and Hoi, S. 2010. Exclusive lasso for multi-task feature selection. In *International Conference on Artificial Intelligence and Statistics*, 988–995.

Zhu, X.; Loy, C. C.; and Gong, S. 2013a. Constrained clustering: Effective constraint propagation with imperfect oracles. In *IEEE International Conference on Data Mining*, 1307–1312.

Zhu, X.; Loy, C. C.; and Gong, S. 2013b. Video synopsis by heterogeneous multi-source correlation. In *IEEE International Conference on Computer Vision*, 81–88.

Zhu, X.; Loy, C. C.; and Gong, S. 2014. Constructing robust affinity graphs for spectral clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1450–1457.

Zhu, X.; Loy, C. C.; and Gong, S. 2015a. Constrained clustering with imperfect oracles. *IEEE Transactions on Neural Networks and Learning Systems* PP(99):1–13.

Zhu, X.; Loy, C. C.; and Gong, S. 2015b. Learning from multiple sources for video summarisation. *International Journal of Computer Vision* 1–22.