Discovering Video Clusters from Visual Features and Noisy Tags

Arash Vahdat, Guang-Tong Zhou, and Greg Mori

School of Computing Science, Simon Fraser University, Canada
{avahdat,gza11,mori}@cs.sfu.ca

Abstract. We present an algorithm for automatically clustering tagged videos. Collections of tagged videos are commonplace, however, it is not trivial to discover video clusters therein. Direct methods that operate on visual features ignore the regularly available, valuable source of tag information. Solely clustering videos on these tags is error-prone since the tags are typically noisy. To address these problems, we develop a structured model that considers the interaction between visual features, video tags and video clusters. We model tags from visual features, and correct noisy tags by checking visual appearance consistency. In the end, videos are clustered from the refined tags as well as the visual features. We learn the clustering through a max-margin framework, and demonstrate empirically that this algorithm can produce more accurate clustering results than baseline methods based on tags or visual features, or both. Further, qualitative results verify that the clustering results can discover sub-categories and more specific instances of a given video category.

1 Introduction

We have witnessed substantial progress in the acquisition and storage of videos. For example, there are 100 hours of videos uploaded to YouTube every single minute [1]. With this rapid increase in the scale of video collection, effective and efficient video analysis techniques are increasingly in demand and crucial for organization of this content.

Automatic clustering of videos is an essential means of video analysis. Clustering has important uses – it can provide users with browsing capability, enabling exploration of the content of a video dataset. It can also provide subcategories, that can for instance be used to train specific detectors for different sub-categories or otherwise provide a more detailed understanding of a topic.

To cluster videos, a straight-forward approach is to extract visual features (e.g. HOG3D [2]) from video appearance, and then apply a standard clustering algorithm. For instance, Wang et al. [3] cluster images strictly based on appearance, and Niebles et al. [4] develop topic models based on video bag-of-words approaches. However, these methods are generally limited in performance due to the lack of semantics in low-level visual appearance.

To bridge the semantic gap, other approaches turn to semantic cues associated with a video. Video tags are often considered for this purpose due to

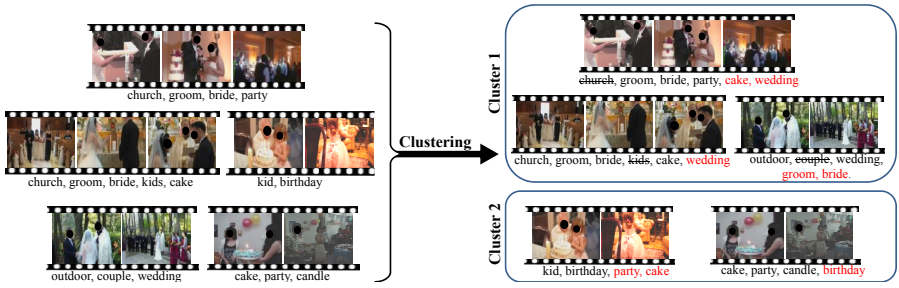

Fig. 1: Video tags provide rich information that can be used for discovering clusters from unconstrained web videos. An algorithm for automatically clustering videos with noisy tags is proposed that explicitly models tag label noise. The proposed approach can be used to remove noisy tags or to add missing tags to a video while finding clusters.

their easy accessibility. These tags range from user-generated content to video captions to semantically meaningful attributes automatically produced by vision algorithms. Different from the low-level visual cues extracted from video appearance, tags provide a complementary view of high-level video semantics. There has been much work in recent years clustering tagged videos. For example, Zeng et al. [5] utilize a learned model over text-based tags to cluster search results. Schroff et al. [6] cluster videos by location based tags. Hsu et al. [7] build hierarchical clustering using tags extracted from user-contributed comments.

Note that visual features and tags are two heterogeneous cues. It is intuitive and beneficial to combine them into a single clustering framework. However, implementing the "combination" is non-trivial. As revealed in our experiments, simply concatenating the two features does not work well since it disregards the hierarchical nature in between low-level visual features and high-level tag semantics. Even a "hierarchical" model – for example, Zhou et al. [8] build tag models from visual features and then use the resultant tag models to assist in clustering – leads to sub-optimal performance. It is because this type of methods learn tag models and clustering in two disjoint processes, and there is no interaction among them. To address the problem, we develop a structured model that jointly considers the interaction between visual features, video tags and video clusters – tags are modeled beyond visual features, and video clusters are determined by jointly examining the tags and visual features.

Another serious problem related to tagged video clustering is that tags are always noisy – obtaining perfect, accurate tags is unlikely in any realistic setting. Consider the examples of tagged videos in Fig. 1. Some tags relevant to the video content are mentioned, while others are "false positives", either irrelevant to the content or not visually discernible. Further, tags are not mentioned with perfect recall – if a tag is not present it does not mean the corresponding concept is not in a video. See the tags highlighted in red in Figure 1 for an example. A recent solution by Vahdat and Mori [9] suggests to "flip" a tag label by revealing the

inconsistency of video appearance with other videos (from the same class) that share the same tag. Note that this solution is tailored to supervised binary classification problems. However, our problem is unsupervised video clustering where no video-level supervision is provided and there are always multiple clusters. To handle tag noise in our settings, we propose to learn cluster-specific models in a joint framework. Each model can actively select high-responding videos as its cluster members. Videos from the same cluster are expected to have similar visual appearance and tags – we allow inconsistent tag labels to flip while penalizing the number of changes.

To implement these ideas, we formulate a max-margin clustering framework that utilizes the conjunction of visual features and explicit models of tag noise. An alternating descent algorithm is developed to effectively solve the resultant non-convex optimization problem. We show empirically that our method aids in clustering, and outperforms approaches that are based strictly on text-based tags or prediction of tags from visual data. Quantitative results show that the method recovers more accurate clusters, and qualitative results further demonstrate the ability to discover sub-categories among videos of a given type.

2 Previous Work

There exists a rich literature on video analysis. In this work, we focus on "unconstrained" videos that are collected by users in a variety of acquisition environments. Specifically, we work on the TRECVID MED collection [10] which provides a standard benchmark for unconstrained web video analysis.

Video analysis mainly includes the tasks of video clustering, video retrieval and video recognition. We have reviewed various video clustering methods above, so we omit their description here for brevity. For video retrieval, significant research effort has been devoted in the form of event detection applied to the TRECVID MED collection. We refer the readers to an excellent state-of-the-art work of Natarajan et al. [11] for more details. Video recognition has been an active research area in computer vision. For example, Izadinia and Shah [12] recognized complex video events (e.g. "parade", "landing fish") from low-level event tags (e.g. "people marching", "person reeling"). A recent work by Vahdat and Mori [9] developed a method for modeling tag label noise for improving video classification. We build on this line of work, instead focusing on the problem of unsupervised video clustering. Furthermore, beyond single event-level labeling for videos, other research has focused on labeling videos with a set of tags. A representative work in this area by Qi et al. [13] predicted multiple correlative tags in a structural SVM framework. The obtained tags can aid in video clustering, retrieval and recognition tasks.

The framework we develop for clustering is based on the max-margin clustering (MMC) approach of Xu et al. [14], which searches for clusterings of input instances that have a large margin between different clusters. As compared to other standard clustering methods (such as K-means and Spectral Clustering), MMC jointly optimizes cluster-specific models and instance-specific labeling as-

signments, and often generates better clusters [14–17]. A recent work by Zhou et al. [8] applied a variant of MMC to cluster videos with latent tags. We extend this framework by two aspects: i) we explicitly model tag noise, and show that incorporating noisy labels can produce more accurate clustering; and ii) (as mentioned before) we use a structured model to capture interaction between visual features, tags and video clusters, instead of using two disjoint processes for learning tag models and clustering [8].

Sub-categories have been studied in the context of fine-grained recognition of a given class or topic. The most popular technique for sub-categorization uses latent variable models: it first utilizes clustering strategies to initialize sub-categories, and then encode sub-category information as latent variables for learning sub-category models. Yang and Toderici [18] use co-watch data to learn sub-categories on YouTube videos. Hoai and Zisserman [19] develop a discriminative approach to sub-category discovery. We show that our clustering algorithm can be used to discover sub-categories within a video category, and different from previous approaches we utilize a structured noisy tag model for this clustering.

3 Tag-based Video Clustering

Consider the problem of discovering clusters of similar videos in unconstrained web videos, similar to YouTube-type videos generated by amateur users. In the most naive way, one can extract visual features from videos and cluster them in visual feature space using off-the-shelf techniques like K-means. The main drawback of this technique is that the formed clusters tend to lack semantic meaning. The problem arises from the underlying visual features used in the clustering. Low level features often fail to represent higher level semantics. Therefore, the resultant clusters are created according to the distance of input samples in the visual feature space which may not match to the conceptual difference that humans associate to the samples.

In contrast, one can explore other sources of information for video clustering rather than pure visual features. Often there are other data available, such as user-provided tags which are common among internet video sharing websites. Tags may refer to objects, actions, scenes or other semantically meaningful entities in a video. Clusters formed on tags are more likely to be semantically meaningful clusters than those created using visual features solely, as the clusters are created in semantically meaningful tag space where similar videos are more likely to share the same tags.

However, tags available on video sharing websites are typically very noisy. The source of noise may vary in different cases, but mainly it can be due to the ambiguity of the process. Tagging is a very subjective task and users may not agree on the tags that should be assigned to the same video. Users can fail to identify some tags relevant to their content; sometimes they introduce spam tags to increase their chance in the retrieval process by misleading the system with tags that are not actually present in their video. In this case, clusters created

from tags will be prone to this noise, and may represent a group of irrelevant videos. On the other hand, obtaining high-quality and noise-free tags can be a very expensive annotation process.

In this paper, we are aiming at an alternative approach to video clustering that works with noisy tags. To implement the idea, we develop a structured model that considers the interaction between video visual features, video tags and video clusters. This structured model enables us to detect tags in a video that are correlated with clusters. In contrast to previous tag-based clustering approaches (e.g. [8]), our model will be equipped with a tag model that recognizes tags on a video using visual features. The tag model will help us detect noise by revealing the inconsistency of a video's visual features with the other samples that share the same tag.

First, we introduce the details of the visual feature-tag-clustering model used for detecting tags and clustering videos. Next, in Sec. 3.2 we present structured max-margin clustering approach followed by our flip max-margin clustering approach that clusters videos using noisy tags in Sec. 3.3. Finally, the details of the optimization are described in Sec. 3.4.

3.1 Cluster Model for Visual Features and Tags

In this work, a structured model is defined for representing the relationship between visual features and tags in a video cluster. The model is designed such that both video clustering and tagging can be performed jointly. For this purpose, we incorporate a tag model to detect tags present in a video, and a tag-cluster interaction model to represent the correlated tags and clusters.

Let us represent a video by x and a set of T binary tags using $\mathbf{t} = \{t_i\}$ for $i = 1, 2, \ldots, T$ where $t_i \in \{-1, 1\}$ represents the presence and absence of i-th tag respectively by 1 and -1. The scoring function $w^\top \phi(x, \mathbf{t}, y)$, which measures the compatibility score between cluster y and tag labeling $\mathbf{t}$ for the video x is defined as:

$$w^\top \phi(x, \mathbf{t}, y) = \sum_{i=1}^{T} t_i \alpha_i^\top \theta(x) + \sum_{i=1}^{T} \beta_{i,y}^\top \varphi(t_i) \tag{1}$$

Here, $\theta(x)$ is a global feature extracted from video x, α_i is the appearance parameter for the i-th label and the term $t_i \alpha_i^\top \theta(x)$ measures the compatibility of the global feature with the i-th tag label. $\varphi(t_i)$ is a vector of size two that indicates whether -1 or 1 has been taken by t_i using $[1, 0]$ or $[0, 1]$, and, $\beta_{i,y}^\top \varphi(t_i)$ measures the compatibility between the i-th tag and the cluster y. Specifically, $\beta_{i,y}$ is a two-dimensional weight vector that represents how likely each case of the tag t_i (e.g. presence as $t_i = 1$ or absence as $t_i = -1$) is associated with the cluster y. Naturally, a large value of $\beta_{i,y}$ on the presence case means that videos in the cluster y tend to have the tag t_i, and a large value of $\beta_{i,y}$ on the absence case means that videos in the cluster y tend to not have the tag t_i. $\{\alpha_i\}_{i=1}^{i=T}$ and $\{\beta_{i,y}\}_{i=1,y=1}^{i=T,y=K}$ are the parameters of our model that are represented altogether by w. Next, the training criterion for learning w is discussed.

3.2 Structured Max-Margin Clustering

The goal of video clustering is to group videos into clusters such that videos in the same cluster are similar. A variety of clustering methods exists in the literature, using different video features and different clustering criteria.

First of all, the features used in clustering have crucial impact on the quality of clusters. Video clustering may be performed over low-level visual features, or semantically meaningful tags, or both.

Apart from the features, the clustering criterion constitutes another dimension of flexibility. Among the widely used approaches, K-means assigns samples to clusters such that intra-cluster variation is minimum, or Spectral Clustering [20] uses eigenvalues of the affinity matrix of the data to map data to an embedding space before clustering. Max-margin clustering (MMC) [14] instead finds a labeling so that the margin between clusters will be maximal.

Here, we extend the MMC approach to the case that there is a structured labeling of tags for each input video available for training. We present a new clustering framework that learns the parameters of the model such that both tag prediction and video clustering can be performed jointly. Given N training videos, $\{x_n, \mathbf{t}_n\}_{n=1}^{n=N}$ to be clustered into K clusters, the goal of structured max-margin clustering (Structured MMC) is to find the labeling $y_n \in \{1, 2, \ldots, K\}$ using the following optimization problem:

$$\begin{aligned} &\min_{w,\xi_n,y_n} \frac{\lambda}{2}||w||_2^2 + \sum_{n=1}^{N} \xi_n \qquad (2)\\ &\text{s.t.} \quad w^\top \phi(x_n, \mathbf{t}_n, y_n) \geq w^\top \phi(x_n, \mathbf{t}, y) + \Delta_{\mathbf{t},\mathbf{t}_n}^{y,y_n} - \xi_n \quad \forall \mathbf{t}, \forall y \\ &\quad L \leq \sum_{n=1}^{N} \mathbb{1}_{(y_n=k)} \leq U \qquad\qquad \forall k \in \{1, 2, \ldots, K\} \end{aligned}$$

which minimizes the norm of parameters $||w||_2^2$ while assigning training examples to clusters as well as tagging them with a minimum structured error measured by the slack variables ξ_n. λ is a hyper parameter that controls the balance between the norm of model parameters and constraint violation. The first constraint enforces that the compatibility score of video x_n, its tag label $\mathbf{t}_n$ and assigned cluster y_n is greater than any other hypothesized labeling. Here, the margin is re-scaled based on how different the hypothesized labeling is from the annotation using the loss function $\Delta_{\mathbf{t},\mathbf{t}_n}^{y,y_n}$. Note that the loss function is a function of both cluster assignments and video tags. Therefore, the Structured MMC defined in Eq. 2 not only maximizes the margin between clusters, but also learns parameters such that the annotated tags have higher scores than hypothesized tag labels.

The second constraint in Eq. 2 enforces balanced clusters where L and U are the lower and upper bounds controlling the size of each cluster. The same constraint is used in [8] to prevent the algorithm from finding the trivial clustering that assigns all the videos to one cluster.

Note that the optimization problem in Eq. 2 differs from previous clustering techniques in that both the structured prediction of tags and clustering are

formulated in a unified framework. This is an essential capability as after learning model parameters, w, using training videos, the model can potentially be used to jointly tag and cluster unseen videos.

3.3 Flip Max-Margin Clustering

The Structured MMC algorithm described above relies on the tags given in the training phase. However, in the case of noisy tags the quality of clusters can be poor since they are formed based on unreliable tags. Further, tags that are missing on training videos can have a significant effect on the clustering results, since the model will unduly penalize their absence on a particular video.

Instead of treating the tags provided on training videos as fixed, we explicitly model the possibility of incorrect tags on input videos. Motivated by the idea of Flip SVM [9], we propose flip max-margin clustering (Flip MMC) that is allowed to change tags in the course of training. In this approach, the training algorithm may correct some tag label noise by considering their inconsistency in visual feature space with respect to the videos sharing the same tag. But at the same time, the algorithm is penalized for label changes to prevent the situation where all the tags are set to the same category.

In order to operationalize this idea, we modify the optimization problem for Structured MMC in Eq. 2. We change this optimization problem to include uncertainty in tags, allowing a certain number of tags to "flip" or change. This will let the clustering algorithm adaptively correct tags on a training video believed to be erroneous, adding missing tags to a video and/or deleting spurious ones.

Let us define the refined tag labels for the n-th training example by $\mathbf{t}'_n = \{t'_{ni}\}_{i=1}^{i=T}$. Intuitively, the refined tag label should be similar to annotated (noisy) tag label, $\mathbf{t}_n$, except a few tags *flipped* based on inconsistency with other videos in the same cluster. Here, the label change cost function $\Delta'_{\mathbf{t}_n,\mathbf{t}'_n}$ is defined to penalize training algorithm from making refined tag label very different from the annotated tags, $\mathbf{t}_n$. In this case, the optimization problem of Flip MMC is formulated as:

$$\begin{aligned}
&\min_{w,\xi_n,\xi'_n,y_n,\mathbf{t}'_n} \frac{\lambda}{2}||w||_2^2 + \sum_{n=1}^{N}\xi_n + \gamma\sum_{n=1}^{N}\xi'_n \qquad (3)\\
&\text{s.t.} \quad \xi'_n \geq \Delta'_{\mathbf{t}_n,\mathbf{t}'_n}\\
&\quad w^\top\phi(x_n,\mathbf{t}'_n,y_n) \geq w^\top\phi(x_n,\mathbf{t},y) + \Delta^{y,y_n}_{\mathbf{t},\mathbf{t}'_n} - \xi_n \quad \forall\mathbf{t},\forall y\\
&\quad L \leq \sum_{n=1}^{N}\mathbb{1}_{(y_n=k)} \leq U \qquad\qquad \forall k\in\{1,2,\ldots,K\}
\end{aligned}$$

which minimizes the norm of parameters $||w||_2^2$ while assigning training examples to clusters as well as recognizing refined tag labels, $\mathbf{t}'_n$ constrained to be similar to annotated tags, $\mathbf{t}'_n$. λ and γ are hyper parameters that controls the balance between the norm of model parameters, constraint violation for refined tag label, ξ_n and the tag label change cost ξ'_n. The second constraint enforces that the

compatibility score of video x_n, its refined tag label $\mathbf{t}'_n$ and assigned cluster y_n is greater than any other hypothesized labeling where the margin is rescaled using $\Delta^{y,y_n}_{\mathbf{t},\mathbf{t}'_n}$ similar to Structured MMC. In this work, both $\Delta'_{\mathbf{t}_n,\mathbf{t}'_n}$ and $\Delta^{y,y_n}_{\mathbf{t},\mathbf{t}'_n}$ are assumed to be a decomposable loss function that can be decomposed to a sum of losses measured on individual tag/cluster annotation. Here, simple hamming loss functions are used for both functions. Note that γ controls the amount of annotated label change allowed in training. By setting it to ∞ Flip MMC becomes Structured MMC.

We emphasize that our framework is general, and can handle video datasets with more or less noisy tags. For example, in a case with less noise, the trade off parameter γ can be set to a large value to prevent too many flips. Or, the label change cost function $\Delta'_{\mathbf{t}_n,\mathbf{t}'_n}$ can be renormalized to penalize flipping erroneous tags less, especially if there is some prior information available regarding the amount of noise for each tag.

3.4 Optimization

The Flip MMC framework proposed in the previous section jointly optimizes the model parameters that describe each cluster, finds the best assignment of videos to clusters, and refines the tag labeling to reduce the noise in tag annotation. Similar to MMC, the Flip MMC optimization is a challenging non-convex optimization problem due to the discrete optimization that assigns videos to clusters and refines tag labels.

Here this non-convex optimization problem is rewritten in unconstrained format as:

$$\min_{w} \frac{\lambda}{2}||w||_2^2 + R_w \tag{4}$$

where R_w is the the risk function defined in the form of an assignment problem:

$$R_w = \min_{y_n} \sum_{n=1}^{N} R'_w(y_n) \tag{5}$$

$$\text{s.t.} \quad L \leq \sum_{n=1}^{N} \mathbb{1}_{(y_n=k)} \leq U$$

where $R'_w(y_n)$ computes the "mis-clustering" cost of assigning the n-th video to the cluster y_n using:

$$R'_w(y_n) = \min_{\mathbf{t}'_n} \max_{y,\mathbf{t}} \left(w^\top \phi(x_n, \mathbf{t}, y) + \Delta^{y,y_n}_{\mathbf{t},\mathbf{t}'_n} - w^\top \phi(x_n, \mathbf{t}'_n, y_n) + \gamma \Delta'_{\mathbf{t}'_n,\mathbf{t}_n} \right). \tag{6}$$

In Eq. 6 annotated tags change to $\mathbf{t}'_n$ such that the error of assigning the video x_n to y_n is minimal while number of changes are being penalized by $\Delta'_{\mathbf{t}'_n,\mathbf{t}_n}$.

In order to address the unconstrained optimization problem in Eq. 4, we develop a coordinate descent-style algorithm shown in the supplementary material. This algorithm alternates between finding the parameters of each cluster (w) and finding an assignment of videos to clusters. The algorithm mainly consists

of three steps performed iteratively. First, "mis-clustering" cost is computed in Eq. 6, and then it is used for computing risk function by solving the assignment problem in Eq. 5. Finally, the model parameters are updated given the risk values using the NRBM approach of Do and Artières [21], which is a non-convex extension of the cutting plane algorithm. The details of the training algorithm can be found in the supplementary material.

4 Experiments

In this section, the proposed video clustering technique is examined for two different tasks. First, the method is evaluated for the general video clustering task based on tags, and later it is used for discovering sub-categories of complex video events.

Dataset: For the experiments, we use our model to cluster web videos in the TRECVID MED 2011 dataset [10]. We use the Event Kit video collection that includes 2379 videos from 15 event categories: "board trick", "feeding animal", "landing fish", "wedding ceremony", and "woodworking project", "birthday party", "changing a tire", "flash mob", "getting a vehicle unstuck", "grooming animal", "making sandwich", "parade", "parkour", "repairing appliance", and "sewing project". Each category contains about 150 videos.

Visual Feature: For all experiments HOG3D features [2], k-means quantized into a 1000-word codebook are used. For all techniques that require visual features, the approximated Histogram Intersection Kernel via feature extension [22] is used to provide higher quality results.

Tags: The noisy tags generated in Vahdat and Mori [9] TRECVID MED 2011 dataset are used in the experiments. [9] uses text analysis tools to extract binary tags based on one-sentence long textual description of videos provided with the dataset in the "judgment files." As tags are generated from arbitrary sentences, there is a large amount of noise inherited in tag annotation. The 114 tags that have more than 10 occurrences in the dataset are used here.

4.1 Video Clustering

In this section, the proposed clustering approach is used to cluster the web videos in the TRECVID MED 2011 dataset. Following previous work [14–17, 8], the videos are grouped into the number of event categories in the dataset ($K = 15$). The Flip MMC and Structured MMC approach are compared with four sets of baselines:

Visual Features: The first set are based on approaches that work directly on the visual features without considering any tag annotation. Here a video is represented by a global bag-of-words feature vector. We have examined three conventional approaches including the K-means algorithm, Spectral Clustering [20], and the MMC approach implemented in [8]. Furthermore, to mitigate the effect of randomness, K-means and Spectral Clustering are run 10 times with different initial seeds and the average results are recorded in the experiments.

Binary Tags: The second set of baselines is the same baselines where visual features are replaced with binary tag annotations. Here a video is represented by a vector of binary variables indicating the annotated presence/absence of tags , and the same K-means algorithm, Spectral Clustering [20], and the MMC approach [8] are used for clustering videos.

Binary Tags and Visual Features: The third set of baselines are created by representing each video using the concatenation of their visual features as well as binary tag labels. The baseline shows the case where information from heterogeneous sources are combined in a naive way.

Detection Scores: The fourth set of baselines trains SVM tag detectors from visual features, and represents each video by a vector of tag detection scores. Note that these baselines consider tag detection and clustering as two separated steps. In contrast, our approach models tags and clusters in a joint framework, while correcting noisy tags. As above, we have conducted K-means, Spectral Clustering and MMC on this data. We have also compared the latent max-margin clustering (Latent MMC) approach proposed in [8], which clusters videos based on the latent presence/absence of video tags. Note that Latent MMC originally builds tag detectors on a different dataset other than the one for clustering. As we assume tag annotation on the clustering dataset, a fair comparison is made by training tag detectors on the same clustering data for all the compared methods.

Parameters: MMC, Latent MMC, Structured MMC and Flip MMC require setting the lower bound (L) and upper bound (U) values in cluster balance constraint. For all these methods we set L and U to $0.9\frac{N}{K}$ and $1.1\frac{N}{K}$ respectively. For all the methods, the trade off parameter, λ is chosen as the best from the range $\{0.1, 1, 10\}$, and the other trade-off parameter of Flip MMC γ, is set to 0.1. All MMC based clustering we used the same initialization of clusters resulted from Spectral Clustering. The same optimization package is used for all the MMC-like methods for a fair comparison.

We use Hamming loss for both Structured MMC and Flip MMC. $\Delta'_{\mathbf{t}',\mathbf{t}}$, the label change cost function is also defined as Hamming loss which basically counts the number of label changes. For flip part, we defined cost function such that it prevents label flips from a positive tag to a negative tag. The rational behind this type of loss function is the fact that in the TRECVID MED dataset, sentences used for generating tag annotation are entered by expert annotators. It is assumed that the annotators have not entered spam sentences. So, the extracted tags are actually present in the video, and there is no need to remove them. However, it is natural to assume that sentences does not contain all the potential tags annotated (mentioned) in the sentence.

Performance measures: Four standard measurements are used to evaluate the quality of the clusters: purity [14] measures the accuracy of the dominating class in each cluster, normalized mutual information (Normalized MI) [23] is from the information-theoretic perspective and calculates the mutual dependence of the predicted clustering and the ground-truth partitions, Rand index [24]

Table 1: Quantitative comparison of clusters generated by different approaches. *Visual Features* represents the set of baselines that perform on visual features, *Binary Tags* are the baselines that work with binary tags directly, *Binary tags and Visual Features* are those that use both visual features and binary tags, and *Detection Scores* denotes the set of baselines that use tag detection scores. *Our model* refers to the models of Structured MMC (defined in Sec. 3.1) and Flip MMC (defined in Sec. 3.3).

	Purity	Normalized MI	Rand index	F-measure
Visual Features:				
K-means	0.26	0.19	0.88	0.14
Spectral Clustering	0.25	0.20	0.88	0.15
MMC	0.25	0.19	0.88	0.14
Binary Tags:				
K-means	0.51	0.52	0.86	0.30
Spectral Clustering	0.71	0.73	0.93	0.56
MMC	0.76	0.72	0.95	0.64
Binary Tags and Visual Features:				
K-means	0.51	0.49	0.90	0.34
Spectral Clustering	0.76	0.74	0.94	0.62
MMC	0.79	0.72	0.95	0.66
Detection Scores:				
K-means	0.63	0.60	0.93	0.50
Spectral Clustering	0.82	0.76	0.96	0.69
MMC	0.83	0.78	0.96	0.73
Latent MMC	0.86	0.82	0.97	0.79
Our model:				
Structured MMC	0.87	0.84	0.97	0.79
Flip MMC	**0.90**	**0.88**	**0.98**	**0.84**

evaluates true positives within clusters and true negatives between clusters and balanced F-measure considers both precision and recall.

Results: The quantitative comparison of the proposed clustering approach with baselines is presented in Table 1. On the TRECVID MED 2011 dataset, Flip MMC achieves the highest performance in terms of all the measurements. The comparison between Structured MMC and Flip MMC shows the efficiency of label flip in getting better clusters. Surprisingly, the performance of K-means, Spectral Clustering and MMC gain a significant boost when discrete tag labels were replaced with the detection scores of an SVM classifier that is trained on the training dataset. This may be due to the fact that SVM maps binary tag labels to a continuous domain where the magnitude of scores are correlated with the strength of the presence of the tag. The comparison between *Visual Features* and *Binary Tags* baseline sets confirms the fact that in general clustering videos based on tags can actually result in semantically meaningful clusters, and finally the low accuracy of *Binary Tags and Visual Features* baselines comparing to

our approach shows that naive approaches such as feature concatenation may improve the accuracy of techniques that rely on individual sources of information such as visual features or tags, but, in contrast our approach can wisely use the information in visual features to refine annotated tag labels that result in better clusters.

4.2 Sub-categorization

In this section, the proposed approach is used to discover sub-categories of event categories in TRECVID MED 2011. Due to the complex nature of events defined in the TRECVID MED dataset, the amount of intra-class variation is very large. Sub-categorization of events can be potentially used to discover clusters that have lower intra-class variation. For example, an event such as "boarding trick" can involve snowboards, skateboards, or other devices; can occur in scenes ranging from urban streets to watery environs; and has other variations, much of which is captured in relevant tags.

In order to discover sub-categories, we consider videos from each ground-truth event category in turn. Videos of each event category are clustered into six clusters. Representative samples of the results for "grooming animal", "getting a vehicle unstuck", and "boarding trick" are visualized in Figure 2. As we are using tags for clustering videos, the discovered clusters are in general semantically meaningful.

Figure 2 shows examples of sub-categories for "getting a vehicle unstuck" that correspond to the type of vehicle, or the environment in which the vehicle has been stuck. Clusters that correspond to getting a vehicle unstuck from mud or snow are discovered. The event category "grooming animal" results in sub-categories that vary according to the animal being groomed, and snowboard/skateboard variants are discovered in the "boarding trick" event category.

5 Conclusion

We have presented a method for automatically obtaining clusters of videos by utilizing visual features and noisy tags. We developed a clustering algorithm based on max-margin clustering that finds groups of videos by optimizing a max-margin criterion separating each cluster from competing ones. Different from previous clustering approaches, we explicitly model label noise. We showed empirically that this was effective, resulting in more accurate clustering than a set of baseline methods.

We presented results on the TRECVID MED unconstrained web video dataset that verified the efficacy of the proposed method. In particular, one could discover either high-level event categories or semantically meaningful sub-categories of events by utilizing noisy tag data in conjunction with visual features. Noisy tag data are commonplace, and methods for effectively using them for clustering could facilitate more efficient methods for exploring and understanding web video collections.

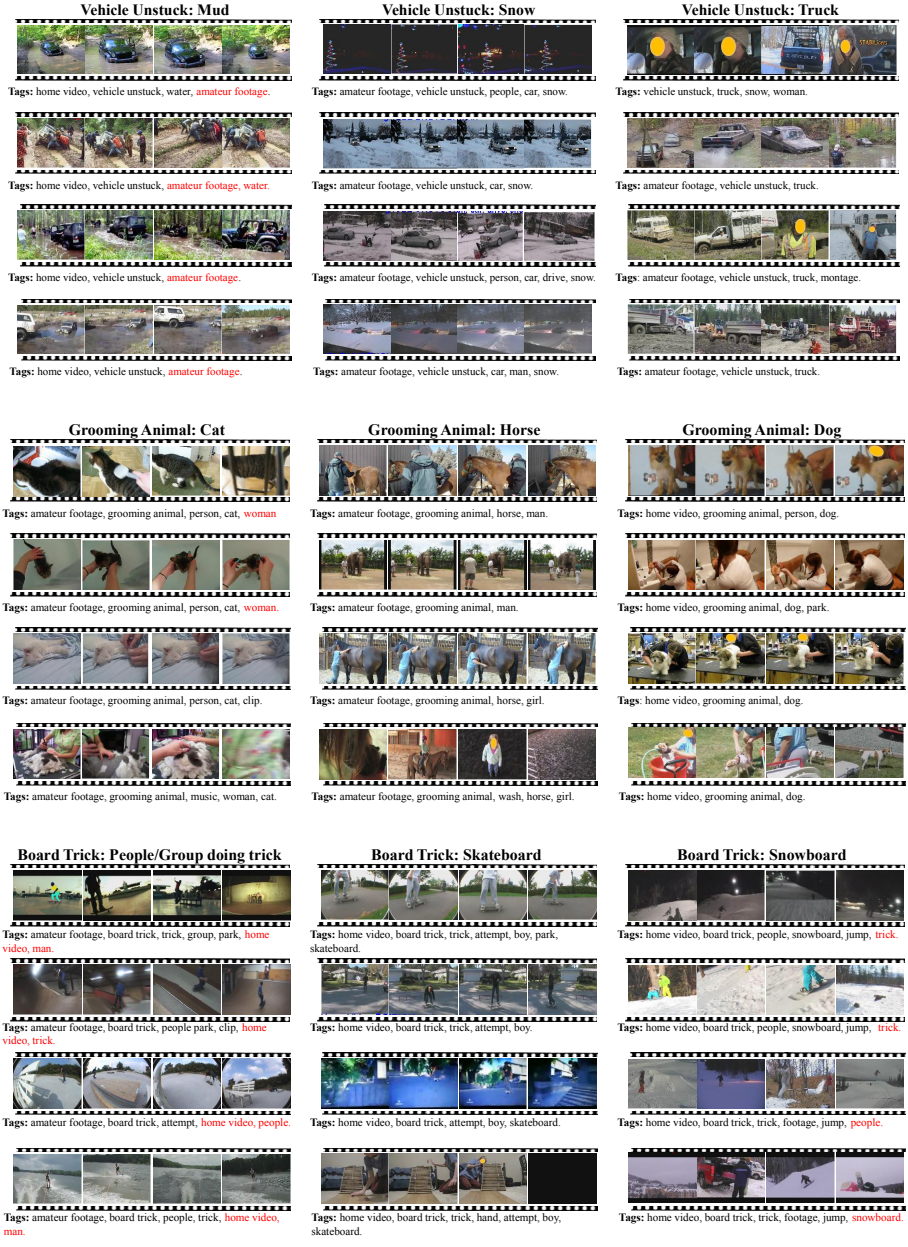

Fig. 2: Qualitative visualization of sub-categories discovered from event categories in TRECVID MED 2011 dataset. Each row represents three sub-categories of an event category. For each sub-category four highest-scored videos are visualized. The tags associated for each video is also reported along with red tags added by Flip MMC. In most cases the formed clusters represent a semantically meaningful sub-category. The semantic content of each cluster can be extracted by manually checking common detected tags, and is reported on top.

References

1. YouTube: Statistics - youtube (2014) [Online; accessed 27-February-2014].
2. Kläser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC. (2008)
3. Wang, Y., Jiang, H., Drew, M.S., Li, Z.N., Mori, G.: Unsupervised discovery of action classes. In: CVPR. (2006)
4. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. In: BMVC. (2006)
5. Zeng, H.J., He, Q.C., Chen, Z., Ma, W.Y.: Learning to cluster search results. In: SIGIR. (2004)
6. Schroff, F., Zitnick, C.L., Baker, S.: Clustering videos by location. In: BMVC. (2009)
7. Hsu, C.F., Caverlee, J., Khabiri, E.: Hierarchical comments-based clustering. In: SAC. (2011)
8. Zhou, G.T., Lan, T., Vahdat, A., Mori, G.: Latent maximum margin clustering. In: NIPS. (2013)
9. Vahdat, A., Mori, G.: Handling uncertain tags in visual recognition. In: ICCV. (2013)
10. Over, P., Awad, G., Michel, M., Fiscus, J., Kraaij, W., Smeaton, A.F., Quenot, G.: TRECVID 2011 — an overview of the goals, tasks, data, evaluation mechansims and metrics. In: TRECVID. (2011)
11. Natarajan, P., Wu, S., Vitaladevuni, S.N.P., Zhuang, X., Tsakalidis, S., Park, U., Prasad, R., Natarajan, P.: Multimodal feature fusion for robust event detection in web videos. In: CVPR. (2012)
12. Izadinia, H., Shah, M.: Recognizing complex events using large margin joint low-level event model. In: ECCV. (2012)
13. Qi, G.J., Hua, X.S., Rui, Y., Tang, J., Mei, T., Zhang, H.J.: Correlative multi-label video annotation. In: ACM MM. (2007)
14. Xu, L., Neufeld, J., Larson, B., Schuurmans, D.: Maximum margin clustering. In: NIPS. (2004)
15. Valizadegan, H., Jin, R.: Generalized maximum margin clustering and unsupervised kernel learning. In: NIPS. (2006)
16. Zhang, K., Tsang, I.W., Kwok, J.T.: Maximum margin clustering made practical. In: ICML. (2007)
17. Zhao, B., Wang, F., Zhang, C.: Efficient multiclass maximum margin clustering. In: ICML. (2008)
18. Yang, W., Toderici, G.: Discriminative tag learning on youtube videos with latent sub-tags. In: CVPR. (2011)
19. Hoai, M., Zisserman, A.: Discriminative sub-categorization. In: CVPR. (2013)
20. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: NIPS. (2001)
21. Do, T.M.T., Artières, T.: Large margin training for hidden markov models with partially observed states. In: ICML. (2009)
22. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. IEEE Trans. on Pattern Analysis and Machine Intelligence **34**(3) (2012) 480–492
23. Kvalseth, T.O.: Entropy and correlation: Some comments. IEEE Transactions on Systems, Man and Cybernetics **17**(3) (1987) 517–519
24. Rand, W.M.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association **66**(336) (1971) 846–850