Improving Semantic Concept Detection through the Dictionary of Visually-distinct Elements

Afshin Dehghan, Haroon Idrees, Mubarak Shah
Center for Research in Computer Vision, University of Central Florida
{adehghan, haroon, shah}@cs.ucf.edu

Abstract

A video captures a sequence and interactions of concepts that can be static, for instance, objects or scenes, or dynamic, such as actions. For large datasets containing hundreds of thousands of images or videos, it is impractical to manually annotate all the concepts, or all the instances of a single concept. However, a dictionary with visually-distinct elements can be created automatically from unlabeled videos which can capture and express the entire dataset. The downside to this machine-discovered dictionary is meaninglessness, i.e., its elements are devoid of semantics and interpretation. In this paper, we present an approach that leverages the strengths of semantic concepts and the machine-discovered DOVE by learning a relationship between them. Since instances of a semantic concept share visual similarity, the proposed approach uses soft-consensus regularization to learn the mapping that enforces instances from each semantic concept to have similar representations. The testing is performed by projecting the query onto the DOVE as well as new representations of semantic concepts from training, with non-negativity and unit summation constraints for probabilistic interpretation. We tested our formulation on TRECVID MED and SIN tasks, and obtained encouraging results.

1. Introduction

For computer vision problems related to video understanding, it has been shown that performance can be increased through the use of intermediate concepts [16, 8, 19], which may belong to any of the four modalities, i.e., audio, image, video or text. In the semantic hierarchy, concepts lie between events and attributes. Events are composed of concepts such as objects, scenes and actions, while attributes refer to the properties possessed or shared by different concepts. Nonetheless, concepts and attributes have been used interchangeably in literature [28]. Concepts have been mostly used as an intermediate representation to improve performance on some other tasks [1]. However, there are other high-level computer vision problems which directly depend on the performance of individual concept detectors, such as, evidence-based reasoning and recounting [4]. In recounting, where the goal is to generate a textual description of contents of a video, the performance depends on the accuracy of each concept detector, as correct temporal ordering of concepts is crucial. Concept detection is also the precursor to localization, as it reduces the search space. Thus, improvement in semantic concept detection, especially when labeled data is not abundant, will improve results of many other dependent problems.

The traditional approach to obtaining such concepts is through manual annotation [14, 5], which is a cumbersome task especially in large datasets containing hundreds of thousands of videos. For such datasets, the annotation has to be restricted to a subset of the data. However, semantic concepts have the benefit of supervised learning since they have labels. To alleviate the issue of manual annotation, there are quite a few recent works that discover 'concepts' automatically and use them to improve detection [26], classification [18], recognition [13, 27] and retrieval [28]. These concepts are learned automatically in an unsupervised fashion from the data. Although, the machine-discovered concepts do not have labels and are not meaningful, they do offer several advantages: 1) they can be learned through unlabeled data, 2) they are expressive of the entire data, and, 3) it is possible to impose certain constraints while discovering these concepts such as separability or orthogonality.

The labeled and unlabeled nature of semantic and machine-discovered concepts, respectively, may suggest one to employ semi-supervised learning to learn better detectors for semantic concepts. But, our experiments using Transductive SVM reveal that this is not the case and machine-discovered concepts result in lower performance for semantic concept detection (see Sec 4). But, since these concepts have been shown to improve object and event detection, they do offer certain advantage that can be exploited to improve the detection of semantic concepts. In this paper, instead of training classifiers on machine-discovered concepts, we select representative element from each machine

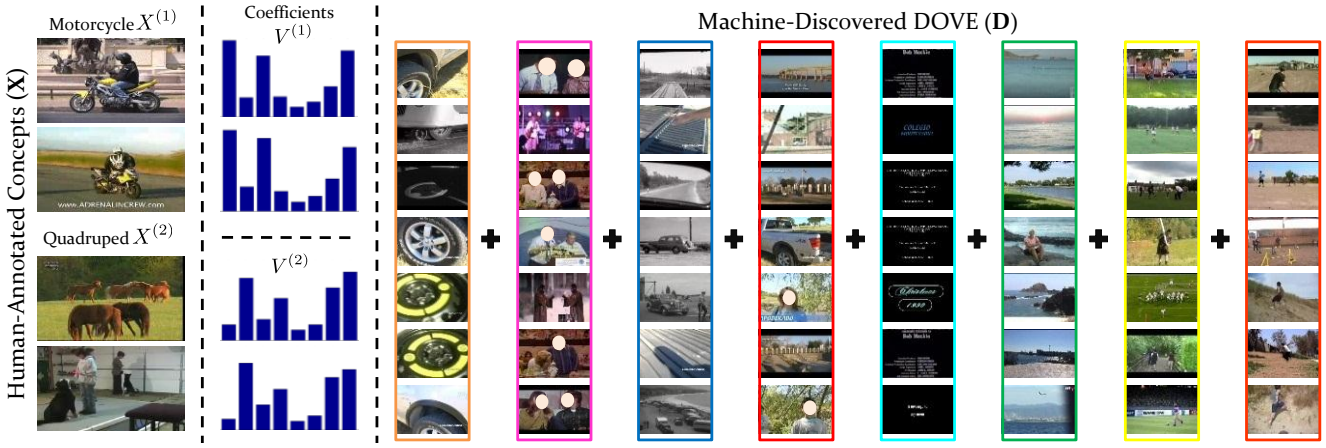

Figure 1: Illustration of the idea: Given pairs of instances from two manually annotated concepts, motorcycle and quadruped, shown in the first column and a dictionary of visually-distinct elements (DOVE) from a given dataset, our goal is to find coefficients for each of the annotated instances in terms of the DOVE. We regularize this using soft-consensus constraint which enforces the new representations to be consistent *across* instances, as can be seen in the pairs of coefficient vectors in the second column.

concept, and form a DOVE. Then, we show that when semantic concepts are described in terms of an over-complete DOVE (Figure 1), it leads to improved performance on semantic concept detection.

Since instances of semantic concepts share visual similarities, we enforce the new representations to be mathematically similar and consistent when describing them in terms of the machine-discovered DOVE. This consistency is achieved through the use of soft-consensus regularizer which captures distance of each instance from the consensus (center) in the new space. Thus, the goal is to reconstruct all the instances of semantic concepts as best as possible while maximizing similarity between the new representations for each individual concept. This procedure yields consistent representations and uses only the positive instances for each concept. We show that training can be performed independently for each semantic concept, which has the advantage that new concepts can be added to the concept pool independently of existing concepts. This characteristic is extremely desirable in large-scale datasets since new concepts can be defined and added without requiring relearning of existing concepts, which will happen for discriminatively-trained detectors due to change in the negative sets. However, our approach does require a complementary testing procedure which takes into account both the DOVE as well as the new representations obtained through training. The proposed approach can be used as an alternative to any dictionary-based classification or detection method [15, 24, 25].

In summary, 1) we explore the idea of improving semantic concepts through machine-discovered DOVE. This is achieved through 2) a training procedure that imposes consistency across samples of each concept. 3) The testing is performed by representing each query in terms of DOVE as well as new representations obtained on semantic concepts from training. 4) The proposed approach is general in nature and can be applied to members of the semantic hierarchy other than concepts, and 5) can supplant existing dictionary-based approaches. 6) The extensive evaluations offer several insights for future research.

2. Related Work

Concepts or attributes have typically been defined manually [14, 12] and are meaningful. Here, we focus on works that use machine-discovered concepts or attributes in isolation or in conjunction with semantic ones. In literature, machine-discovered is synonymous with data-driven [26], latent [6], weak [28], and automatically-discovered [2].

Some works aim to discover attributes that have weak semantics such as Classemes [20] and automatic attributes [2], while others do not care about semantics at all [17, 26, 27]. Both [26, 27] discover machine concepts for complex event detection. Kumar et al. [11] propose to use semantic attributes with simile classifier to improve face verification. Liu et al. [13] also combine semantic and machine-discovered attributes for action recognition. The weight of both type of attributes for each class is learned through Latent SVM. Yu et al. [28] propose to use machine-discovered attributes for image retrieval where the attributes of a query are represented in terms of a small subset of large pool of weak attributes. However, in all these methods, the goal is to improve category (object, event) classification and retrieval. In this work, we propose to improve detection of semantic concepts through a novel approach which utilizes DOVE created from machine-discovered concepts.

Dictionary-based Classification has been extensively used in computer vision especially in face recognition. The goal is to recognize a face image from a training set which may contain multiple face images per person. The feature vectors from training images are directly used as dictionary and the query face image is represented in terms of this dictionary. The query is assigned the identity of the person which has the lowest reconstruction error. Wright et al.[24] introduced sparse representation-based classification (SRC) for face recognition where a single query image is represented as sparse combination of training images. Linear Regression Classification (LRC) was proposed by [15] who used least-squares estimation for classification. Recently, Zhang et al. [29] showed that the success of classification methods based on sparse representation is largely due to collaborative representation, i.e., expressing query image in terms of dictionary of training images, rather than sparsity. The authors propose to use ℓ_2-norm which they show yields similar results as SRC. Since sparsity is useful for compression, and does not have significant influence on classification [29], our experiments for concepts also yield the same conclusion. Note that detection is closely related to classification. While classification picks the class with lowest error for dictionary-based methods, the error can be used as score for detection, where low error means higher likelihood of presence of object, action or event. Although there has been extensive work on learning dictionaries, however, like the papers discussed in this category, we assume that the dictionary is given in advance, which in our case, is obtained *independently* from unlabeled data.

Simultaneous Representation of multiple vectors was independently introduced as Multiple Measurement Vectors (MMV) by Cotter et al. [3] and Simultaneous Sparse Approximation by Tropp et al. [22, 21]. MMV occur in many applications such as neuromagnetic imaging where it is assumed that the input vectors share a common sparsity structure. The aim is then to find sparse approximation of several input vectors simultaneously using different linear combinations of the same base vectors (elements of dictionary). An extension of FOCUSS - an algorithm for finding sparse representation of single input vector - to multiple vector was presented in [3]. Since then several extensions of existing sparse approximation [23] and new algorithms with temporal extensions [29] have been presented. Again, a complete review is beyond the scope of this work. The proposed approach for training differs from MMV in that we do not impose sparsity on the coefficient vectors and a *soft consensus* penalty is imposed on coefficient vectors of instances belonging to the same concept / class. Consensus regularization has also been used recently in [25] to yield consistent coefficient vectors of the same query across different views / features. In contrast, we propose to use it across training instances of a particular semantic concept.

3. Proposed Approach

Our approach begins by discovering machine concepts automatically from a given set of images, keyframes or videos, from which we form the dictionary of visually-distinct elements (DOVE). Next, using the training data, we learn new representations that are consistent across each concept. Then, given a query image or video, we reconstruct it in terms of both the DOVE and new representations from the training data such that coefficients are non-negative and have unit summation. The sum of coefficients for each training concept gives the probability or score of query for belonging to that concept.

3.1. Creating DOVE

To create DOVE, we first discover machine concepts using an approach similar to [26]. Given training data, each image or clip from a video is represented using feature vectors, which can be Bag-of-Words representation based on low-level features. Since the features tend to lie in a high-dimensional non-Euclidean space, a low-dimensional representation is learned that preserves non-linear relationship between the data points. For that we use Deep Belief Network (DBN) by stacking several layers of Restricted Boltzmann Machines (RBMs), where the output from the layer below is used as input to the layer above, with original feature vectors as input to the bottom layer. Each layer above has fewer number of nodes than the one below to reduce the dimensions. Once the network is trained, all images or clips are passed through the network to obtain the new representation. We then cluster them into d groups, and select the medoids as our DOVE. Each machine concept provides an element in the dictionary $\mathbf{D}$ to which the relationship is learned from the semantic concepts. The vectors for semantic concepts are also passed through the same DBN to obtain vector representation that lie in the same space as that of DOVE. Note that, we followed this approach due to its simplicity as sophisticated methods can be used to create DOVE.

3.2. Representing Semantic Concepts in terms of DOVE (Training)

Let C be the total number of concepts (classes) and the training vectors for concept c be given by $X^{(c)} \in \mathbb{R}^{m \times n_c}$. Then the matrix $\mathbf{X} = [X^{(1)}\ X^{(2)}\ \ldots\ X^{(C)}]$ contains training vectors from all concepts. Given the DOVE $\mathbf{D} \in \mathbb{R}^{m \times d}$, we minimize the following objective function:

$$L = \|\mathbf{X} - \mathbf{DV}\|_F^2 + \mu\|\mathbf{V} - \tilde{\mathbf{V}}\|_F^2$$
$$= \mathrm{Tr}\Big((\mathbf{X} - \mathbf{DV})(\mathbf{X} - \mathbf{DV})^T\Big) + \mu\,\mathrm{Tr}\Big((\mathbf{V} - \tilde{\mathbf{V}})(\mathbf{V} - \tilde{\mathbf{V}})^T\Big),$$

where the goal is to obtain coefficient vectors, $V^{(c)} \in \mathbb{R}^{d \times n_c}$, which serves as the mapping from semantic con-

cept c to the DOVE. The coefficient vectors for all concepts are given by $\mathbf{V} = [V^{(1)}\ V^{(2)}\ \dots\ V^{(C)}]$. The first term in above equations capture the reconstruction cost while the second term is the soft-consensus regularizer which enforces the solution to be consistent or similar across vectors of each concept. Mathematically, this also prevents overfitting or for the under-constrained case (large d), it permits the solution to be computed. The matrix $\tilde{\mathbf{V}}$ contains consensus vectors from all concepts. The consensus vector for concept c when replicated n_c times gives $\tilde{V}^{(c)} \in \mathbb{R}^{d \times n_c}$. Thus, $\tilde{\mathbf{V}} = [\tilde{V}^{(1)}\ \tilde{V}^{(2)} \dots\ \tilde{V}^{(C)}]$.

The trace of a matrix does not change by altering off-diagonal entries. Substituting the original matrix with a block-diagonal matrix, where each block belongs to one concept, we can write

$$L = \sum_{c=1}^{C} \mathrm{Tr}\Big((X^{(c)} - \mathbf{D}V^{(c)})(X^{(c)} - \mathbf{D}V^{(c)})^T\Big) + \mu \sum_{c=1}^{C} \mathrm{Tr}\Big((V^{(c)} - \tilde{V}^{(c)})(V^{(c)} - \tilde{V}^{(c)})^T\Big) \quad (1)$$

It is obvious from above equation that $\partial L / \partial V^{(c)}$ does not depend on either $X^{(c')}$ or $V^{(c')}$ when $c' \neq c$. Thus, we can find new representations $V^{(c)}$ for each concept independent of other concepts. In the following, we drop the index of concept for clarity. For a particular concept c, we have

$$L^{(c)} = \mathrm{Tr}\Big((X - \mathbf{D}V)(X - \mathbf{D}V)^T\Big) + \mu\, \mathrm{Tr}\Big((V - \tilde{V})(V - \tilde{V})^T\Big) \quad (2)$$

The matrix $\tilde{V}$ contains repetitions of consensus vector obtained by minimizing (2) w.r.t $\tilde{V}$, i.e., $\bar{V}^{(c)} = n^{-1} V \mathbf{1}_{n_c}$. Substituting $\tilde{V}$ in (2) and expanding the relevant terms,

$$L^{(c)} = \mathrm{Tr}(-2V^T\mathbf{D}^T X + V^T\mathbf{D}^T\mathbf{D}V) + \mu\, \mathrm{Tr}(VJJ^TV^T), \quad (3)$$

where $J = I_{n_c} - n_c^{-1}\mathbf{1}_{n_c}\mathbf{1}_{n_c}^T$ is the centering matrix. Minimizing (3) by taking its derivative w.r.t V, we have

$$\frac{\partial L^{(c)}}{\partial V} = -2\mathbf{D}^T X + 2\mathbf{D}^T\mathbf{D}V + \mu 2VJJ^T, \quad (4)$$

Since J is a projection matrix, we have $J = JJ^T$. Setting (4) equal to zero, we get

$$-\mathbf{D}^T X + \mathbf{D}^T\mathbf{D}V + \mu VJ = 0, \quad (5)$$

$$-\mathbf{D}^T X + \mathbf{D}^T\mathbf{D}V + \mu V I_{n_c} - \mu n_c^{-1} V \mathbf{1}_{n_c}\mathbf{1}_{n_c}^T = 0, \quad (6)$$

Using the fact that $I_d V = V I_{n_c}$, we get the following multiplicative update rule for iterations:

$$V \leftarrow (\mathbf{D}^T\mathbf{D} + \mu I_d)^{-1}\Big(\mathbf{D}^T X + \mu n_c^{-1} V \mathbf{1}_{n_c}\mathbf{1}_{n_c}^T\Big). \quad (7)$$

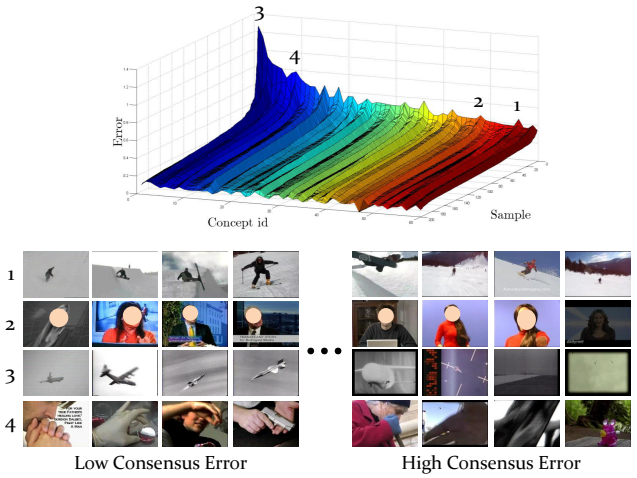

Figure 2: Effect of consensus on training: Since (2) attempts to minimize the distance of each vector to consensus vector, the V of each instance will have a different distance from the corresponding $\bar{V}$. In the figure, we visualize the instances based on their distance from the consensus vectors. On top, we show consensus errors in sorted fashion for all 60 concepts. On bottom, instances are shown based on the error for four concepts (ski, anchorperson, airplane, hand). It is evident that instances of concepts with low training error (1 & 2) are visually more similar. Furthermore, it can be seem that instances with low consensus errors are good exemplars of their respective concepts.

Using this approach, V typically converges in $3-4$ iterations. Figure 2 shows the effect of this procedure on instances from different concepts.

3.3. Concept Detection (Testing)

The training procedure gives new representations for each training instance, therefore, in testing, we have to find new representation for a query both in terms of the DOVE and new representations of training data. During testing, we represent each query instance in terms of the DOVE, and enforce the constraint that the coefficient vector of the query should lie in the space spanned by training coefficient vectors. Thus, the testing becomes a problem of representing query in terms of two dictionaries, i.e., the DOVE and the other comprising training coefficient vectors. We show that this optimization can be solved using quadratic programming which yields a global optimum for each query. Additional constraints such as non-negativity and unit summation are imposed for probabilistic interpretation.

Given a particular query vector $\mathbf{x} \in \mathbb{R}^m$, and all coefficient vectors obtained after training $\mathbf{V} \in \mathbb{R}^{d \times n}$ where $n = \sum n_c$, our aim is to obtain to reconstruct $\mathbf{x}$ in terms of $\mathbf{D}$ while enforcing the constraint that the resultant coefficient vectors lies in the convex hull of $\mathbf{V}$. We pose the

optimization as following:

$$
\begin{aligned}
T &= 2^{-1}\|\mathbf{x} - \mathbf{D}\mathbf{v}\|_2^2 \text{ s.t. } \mathbf{v} = \mathbf{V}\mathbf{y},\ \|\mathbf{y}\|_1 = 1,\ \mathbf{y} > 0 \\
&= 2^{-1}\|\mathbf{x} - \mathbf{D}\mathbf{V}\mathbf{y}\|_2^2 \quad \text{s.t.} \quad \|\mathbf{y}\|_1 = 1,\ \ \mathbf{y} > 0 \\
&= 2^{-1}(\mathbf{x}^T\mathbf{x} - 2\mathbf{y}^T\mathbf{V}^T\mathbf{D}^T\mathbf{x} + 2\mathbf{y}^T\mathbf{V}^T\mathbf{D}^T\mathbf{D}\mathbf{V}\mathbf{y}) \\
&\text{s.t. } \|\mathbf{y}\|_1 = 1,\ \ \mathbf{y} > 0, \qquad (8)
\end{aligned}
$$

which can be solved using quadratic programming:

$$\min_{Y}\ \mathbf{y}^T Q\mathbf{y} + R^T\mathbf{y}\ \text{ s.t. }\ -I_n\mathbf{y} \leq 0\ \ \mathbf{1}_n^T\mathbf{y} = 1, \qquad (9)$$

where $Q = \mathbf{V}^T\mathbf{D}^T\mathbf{D}\,\mathbf{V}$ and $R = -\mathbf{V}^T\mathbf{D}^T\mathbf{x}$. Then the score or probability of query image or video represented by $\mathbf{x}$ belonging to a concept c is obtained by adding the coefficients corresponding to a particular concept.

$$\text{score}(c) = \{\sum_i y_i | \mathbf{V}_{:,i} \in V^{(c)}\}. \qquad (10)$$

Under non-negativity, the unit summation constraint becomes $|\mathbf{y}|_1 = 1$, which means the proposed testing also imposes a weak notion of sparsity. However, since our goal is detection or classification and not sparsity, these constraints come with all the benefits of sparsity (Sec. 4), but allow a global and efficient solution to be computed in the same amount of time as sparse representation. In Figure 3, we show the results of testing procedure in relation to training.

4. Experiments

We performed extensive evaluations of proposed method on both static concepts in images and action concepts in videos using the challenging multimedia datasets of TRECVID SIN and MED. In TRECVID Semantic Indexing (SIN) task, given a set of static training images for different concepts, the goal is to detect concepts that occur in the keyframes of a video. The task is challenging due to low resolution of the images and huge intra-class variation. TRECVID MED is an extremely challenging video dataset with large camera motions, cluttered background and changes in illumination. Furthermore, the dataset is characterized by videos of varying lengths, ranging from few seconds to few minutes, the frame rate lies between 12 to 30 fps and the resolution ranges from 320×480 to 1280×2000. For experiments related to action concepts, we used concepts selected from TRECVID MED data similar to [9]. We report the performance using mean average precision (mAP) [5].

SIN Concepts. For the TRECVID 2013 semantic indexing task, 60 concepts were defined. The set contains variety of concepts including scenes, e.g., *classroom*, *hills*, *lakes*, objects, e.g., *airplane*, *telephone* and activities, e.g., *cheering*, *running*. We used the same set of concepts in our experiments but limit the number of examples per concept to only 200. The images were represented in terms of ISA [7] features with a codebook size of $1,000$. The dimension

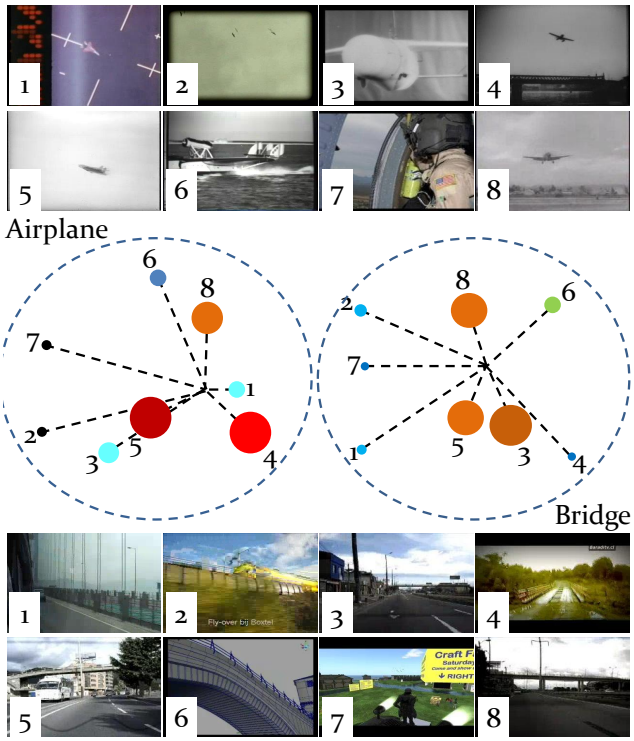

Figure 3: Consensus error in training vs. detection score in testing: For two concepts (*airplane* and *bridge*), consensus error is visualized with dashed lines where longer lines represent higher consensus error during training. Both the color and size of the circles indicates the testing score the instances received (in different folds). This shows that testing score is inversely related to consensus error during training, i.e the length of the dashed lines is inversely proportional to the area of the corresponding circle.

was reduced to 100 using DBN of depth 3 while the DOVE was created from a separate set with $d = 2,000$.

MED Action Concepts. Similar to [9], we manually selected 52 action concepts, e.g, *person dancing*, *person jumping*, *blowing candle*, from the description provided in the event collection (EC) of MED 2011 and 2012. For each concept, we annotated clips by tagging the beginning and ending frame of the video that contains that particular concept. We obtained a total of $9,052$ training clips for the 52 manually defined concepts. Each annotated clip was represented in terms of ISA feature using the codebook size of $4,096$. The dimension was reduced to 500 using DBN of depth 3. We used a subset of clips in the test set DEVT to create DOVE, thus, making sure that there was no overlap between the clips used for DOVE and the testing data. In total, $2,500$ machine-discovered concepts were used to create DOVE.

For both SIN and MED Action concepts, we evaluate the performance by changing the number of training sam-

	Method	K					
		5	25	45	65	85	100
SIN	SVM	11.7 ± 1.5	21.4 ± 1.1	24.1 ± 0.8	25.5 ± 0.8	26.5 ± 0.9	27.3 ± 1.0
	TSVM	08.3 ± 0.6	11.9 ± 1.0	13.6 ± 0.7	15.1 ± 0.6	15.9 ± 0.8	16.8 ± 0.6
	LRC	15.3 ± 1.3	16.1 ± 0.7	15.9 ± 0.5	15.6 ± 0.4	15.3 ± 0.4	07.5 ± 0.6
	SRC	17.4 ± 1.2	24.2 ± 1.1	26.2 ± 0.8	27.1 ± 0.8	27.8 ± 0.8	28.1 ± 0.8
	CRC	11.9 ± 0.7	20.3 ± 0.9	23.0 ± 0.7	24.6 ± 0.7	25.4 ± 0.6	26.0 ± 0.6
	Proposed	**20.0 ± 0.8**	**28.7 ± 0.5**	**32.4 ± 0.3**	**35.6 ± 0.5**	**37.2 ± 1.1**	**38.1 ± 0.3**
MED	SVM	15.5 ± 1.3	25.7 ± 1.2	29.0 ± 1.2	30.6 ± 1.2	31.8 ± 0.5	32.6 ± 0.8
	TSVM	08.1 ± 0.7	13.6 ± 0.3	15.2 ± 0.6	19.5 ± 1.4	21.8 ± 1.4	22.7 ± 0.9
	LRC	19.9 ± 0.6	24.7 ± 0.9	26.2 ± 0.6	26.9 ± 0.9	28.0 ± 0.4	28.0 ± 0.4
	SRC	18.9 ± 1.0	28.5 ± 0.3	34.6 ± 1.3	36.6 ± 0.6	40.0 ± 0.5	39.9 ± 0.5
	CRC	19.9 ± 1.4	24.8 ± 0.9	30.5 ± 1.0	32.2 ± 1.1	34.3 ± 0.4	35.5 ± 1.1
	Proposed	**21.5 ± 1.0**	**33.2 ± 0.5**	**38.9 ± 1.4**	**42.8 ± 0.3**	**45.6 ± 1.1**	**46.6 ± 0.5**

Table 1: Comparison: This table shows the evaluation using SVM, TSVM [10], LRC [15], SRC [24], CRC [25] and the proposed approach which outperforms all other methods.

ples from $K = 5$ to $K = 100$. The experiments were performed on 25 randomly selected concepts and were repeated 10 times for each method. The final mAP as well as the standard deviation is reported in Tables 1 and 2.

Comparison. We compare proposed approach to several baseline methods including SVM, Transductive SVM [10], LRC [15], SRC [24] and CRC [25]. These results are shown in Table 1 for both SIN and MED Action concept. Except for LRC, all methods show an increase in performance as the number of training instances varies from $K = 5$ to $K = 100$. The reason for drop in performance of LRC may be due separate projections onto training classes unlike all other methods which simultaneously project query onto training instances from all classes. The performance of TSVM is much lower than that of regular SVM because machine-discovered concepts add noise in the learning stage and cannot be treated as unlabeled data for semi-supervised learning. The reason is obvious since machine concepts do not have unique labels in terms of semantic concepts and there is one-many and many-one relationship between them. CRC which uses ℓ_2 performs slightly worse than SRC which uses ℓ_1, with a difference of less than 2% on average. The proposed approach gives improved performance compared to all baseline methods. The contribution of different aspects of the proposed approach are given in Fig. 4.

Cross-Combinations. If we were to test the proposed idea of learning the relationship between semantic and machine-discovered DOVE using existing tools, one solution is to use Multiple Measurement Vectors (MMV) for training and SRC for testing. We show the result of MMV/SRC combination in the first two rows of Table 2.

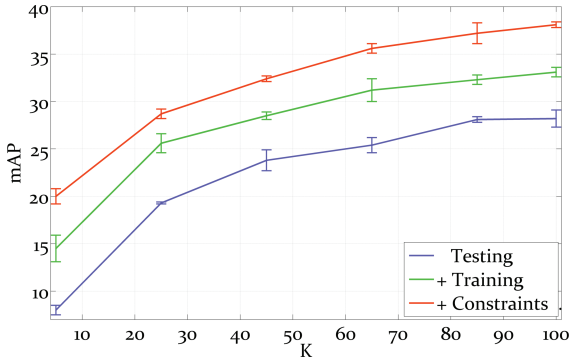

Figure 4: Evaluation of Contributions: The blue curve shows mAP using proposed testing when training instances are used as dictionary and constraints in 8 are ignored. The green curve shows improvement from proposed training using machine-discovered DOVE, while red curve shows the results of proposed approach.

The difference between mAP of proposed approach (Table 1 last row) and MMV/SRC increases from 1% at $K = 5$ 10% at $K = 100$. We tested MMV using two algorithms, MFOCUSS [3] and MSBL [23]. Next, we substitute proposed testing with the two algorithms for MMV. The MMV/Proposed combination performs 3% lower than combination of proposed training and testing. The next five rows show results of using proposed training with LRC, CRC and SRC, Group Lasso and Sparse Group Lasso. Group Lasso performs surprisingly well at all ranges. We used ℓ_2 within groups, which promotes non-sparsity, while ℓ_1 across groups which promotes sparsity. This results in either all instances of a concept selected, with non-zero

Method	K					
	5	25	45	65	85	100
MFOCUSS / SRC	17.8 ± 1.2	25.1 ± 1.2	27.0 ± 0.8	27.3 ± 1.3	29.6 ± 0.6	28.7 ± 0.7
MSBL / SRC	16.7 ± 1.7	24.8 ± 0.4	26.8 ± 0.3	26.6 ± 0.9	27.1 ± 0.7	28.2 ± 1.8
MFOCUSS / Proposed	18.3 ± 0.6	26.2 ± 0.7	30.2 ± 0.7	33.8 ± 0.4	34.3 ± 0.4	34.1 ± 0.6
MSBL / Proposed	17.9 ± 1.5	27.7 ± 0.5	31.2 ± 0.6	33.3 ± 0.6	34.0 ± 1.1	34.1 ± 0.9
Proposed / LRC	15.2 ± 0.5	16.0 ± 0.5	16.3 ± 0.4	16.0 ± 0.6	16.1 ± 0.5	8.6 ± 0.6
Proposed / SRC	17.7 ± 0.9	24.3 ± 1.7	26.2 ± 0.7	27.8 ± 1.4	29.0 ± 0.6	29.3 ± 1.1
Proposed / CRC	10.8 ± 1.4	20.6 ± 0.6	23.2 ± 1.0	23.7 ± 0.5	25.4 ± 0.5	26.1 ± 0.6
Proposed / Lasso ($\ell_1\ell_2$)	17.4 ± 0.5	27.2 ± 1.1	30.3 ± 0.3	31.3 ± 0.6	32.7 ± 0.5	33.9 ± 1.4
Proposed / Lasso ($\ell_1\ell_2$+ℓ_1)	17.3 ± 1.3	28.2 ± 0.7	30.2 ± 1.3	31.9 ± 0.7	32.8 ± 0.4	34.9 ± 0.6

Table 2: Cross-Combinations of training and testing on SIN Concepts: In this table, we report results of different combinations of training and testing. We used MFOCUSS [3] and MSBL [23] for MMV with SRC and proposed testing which outperforms the MMV/SRC by 6%. Next, the combinations of proposed training and LRC [15], SRC [24], CRC [25], Group Lasso and Sparse Group Lasso with ℓ_2-norm is reported. The improvement from proposed training and testing is evident from these evaluations.

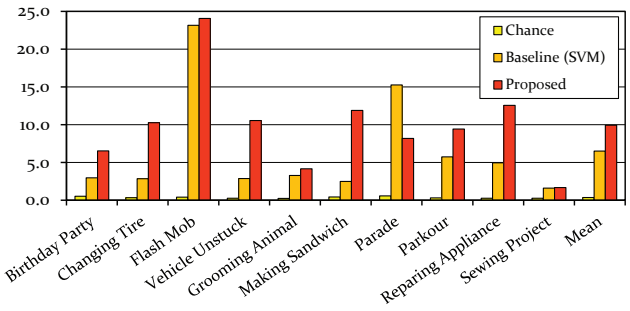

Figure 5: Results on MED Videos: The results of chance, SVM and proposed method are shown in this figure. On average, the proposed approach improves results by 3.2%, a relative improvement of 52%.

coefficients, or all rejected. Sparse Group Lasso differs from Group Lasso with ℓ_1 on all coefficients in addition to group constraints. The results are slightly better than regular Lasso at higher values of K, similar to the difference between CRC and SRC. This suggests that for the task of detection or classification, group sparsity is more important than the sparsity imposed uniformly on all coefficients. The proposed method, despite the fact it does not use group sparsity in testing, outperforms all baseline methods by a margin of 4%. However, we believe when incorporated within our testing, group sparsity will improve the performance of our method as well.

Complex Event Detection. Although this work primarily works on static or video concepts, it is equally applicable to objects or events. We also evaluated our method on TRECVID multimedia event detection (MED) 2011 task where instead of concepts, where we used labeled event videos in place of semantic concepts and videos from hold-out unlabeled data to obtain machine-discovered events which are used to create DOVE. The dataset consists of consumer videos collected from YouTube. For training, there are 10 pre-defined events with $\sim$ 150 videos per event, while for testing, there are $32,061$ videos which may belong to one of the pre-defined events from training set or to none at all. These are termed background videos or null events, and their numerous number significantly affects the average precision, making it a challenging task.

The results of event detection are shown in Fig. 5. The proposed approach gives a relative improvement of 52% over SVM. Performance on all events improves over the baseline except the Parade event. This is due to its visual similarity with Flash Mob, which also depicts crowded scenes. Since we do not learn any discriminative classifiers, it confuses both events thus resulting in lower gains on Flash Mob event and reduced performance in Parade event. The advantages of generative approach for large-scale datasets are discussed in Sec. 1.

Effect of d. The effect of dictionary size, d, is shown in Fig. 6 evaluated at $K = 25$. The x-axis ranges from 100 to 3000, while y-axis shows mAP. The curves for both SIN and MED concepts reach a plateau at $d = 2500$.

5. Conclusion

We explored the problem of improving concept detection by representing semantic concepts in terms of a dictionary of visually-distinct elements created from concepts automatically discovered by machine. We showed that a dictionary-based formulation leads to an improvement in performance of semantic concepts. The proposed training

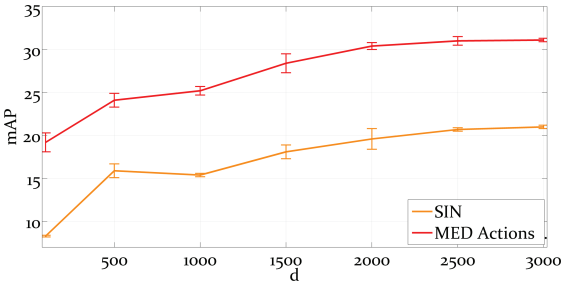

Figure 6: On x-axis is the dictionary size, and on y-axis is mAP. The experiment was performed at $K = 25$.

can be used in place of Multiple Measurement Vectors when sparsity is not desired. An important conclusion is that consistency whether enforced as soft-consensus in training or grouping in testing significantly improves results. Future work includes kernelization of proposed approach and enforcing consensus (grouping) in testing. Using other methods to obtain DOVE which have useful mathematical properties such as orthogonality will also be explored.

Acknowledgments This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20066. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S.Government. We also thank Dr. Xin Li for the exchange of ideas that helped improve this article.

References

[1] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra. Event detection and recognition for semantic annotation of video. *Multimedia Tools and Applications*, 2011.

[2] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010.

[3] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado. Sparse solutions to linear inverse problems with multiple measurement vectors. *Signal Processing*, 2005.

[4] D. Ding et al. Beyond audio and video retrieval: towards multimedia summarization. In *ACM Multimedia Retrieval*, 2012.

[5] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *IJCV*, 2010.

[6] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Attribute learning for understanding unstructured social activity. In *ECCV*. 2012.

[7] A. Hyvarinen and P. Hoyer. Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. In *Neural Computation*, 2000.

[8] N. Inoue, Y. Kamishima, K. Mori, and K. Shinoda. Tokyotechcanon at TRECVID 2012. In *NIST TRECVID Workshop*, 2012.

[9] H. Izadinia and M. Shah. Recognizing complex events using large margin joint low-level event model. In *ECCV*, 2012.

[10] T. Joachims. SVMLight: Support vector machine. *University of Dortmund, http://svmlight. joachims. org/*, 1999.

[11] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *CVPR*, 2009.

[12] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, 2010.

[13] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, 2011.

[14] N. Naphade et al. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 2006.

[15] I. Naseem, R. Togneri, and M. Bennamoun. Linear regression for face recognition. *PAMI*, 2010.

[16] P. Natarajan et al. BBN viser TRECVID 2011 multimedia event detection system. In *NIST TRECVID Workshop*, 2011.

[17] M. Rastegari, A. Farhadi, and D. Forsyth. Attribute discovery via predictable discriminative binary codes. In *ECCV*, 2012.

[18] V. Sharmanska, N. Quadrianto, and C. H. Lampert. Augmented attribute representations. In *ECCV*, 2012.

[19] C. Snoek et al. The MediaMill TRECVID 2012 semantic video search engine. In *NIST TRECVID Workshop*, 2012.

[20] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *ECCV*, 2010.

[21] J. A. Tropp. Algorithms for simultaneous sparse approximation. Part II: Convex relaxation. *Signal Processing*, 2006.

[22] J. A. Tropp, A. C. Gilbert, and M. J. Strauss. Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. *Signal Processing*, 2006.

[23] D. P. Wipf and B. D. Rao. An empirical bayesian strategy for solving the simultaneous sparse approximation problem. *Signal Processing*, 2007.

[24] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *PAMI*, 2009.

[25] M. Yang, L. Zhang, D. Zhang, and S. Wang. Relaxed collaborative representation for pattern classification. In *CVPR*, 2012.

[26] Y. Yang and M. Shah. Complex events detection using data-driven concepts. In *ECCV*, 2012.

[27] F. Yu, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang. Designing category-level attributes for discriminative visual recognition. 2013.

[28] F. X. Yu, R. Ji, M.-H. Tsai, G. Ye, and S.-F. Chang. Weak attributes for large-scale image retrieval. In *CVPR*, 2012.

[29] Z. Zhang and B. D. Rao. Sparse signal recovery with temporally correlated source vectors using sparse bayesian learning. *J-STSP*, 2011.