Toward Improving The Visual Characterization of Sport Activities With Abstracted Scene Graphs

Amir M. Rahimi amrahimi@hrl.com
Kevin Lee klee1@hrl.com
Amit Agarwal aagarwal@hrl.com
Hyukseong Kwon hkwon@hrl.com
Rajan Bhattacharyya rbhattac@hrl.com
HRL Laboratories, 3011 Malibu Canyon Road, Malibu, CA, 90265

Abstract

We present techniques for abstracting relevant information from scene graph features to improve action recognition in sports videos. Feature representation with relevant information can dramatically increase machine learning's utility across many tasks. Despite the advantages of incorporating objects and relations as building blocks of semantic information, we still encounter too many irrelevant objects and relations in sports videos, adding uncertainty to the classifiers. This paper describes four fundamentally different scene abstraction techniques, each searching for the relevant information within aggregated features from pixel-level to object-level. In each method, we formulate relevancy through co-occurrence statistics, semantic similarity, feature decomposition, and correlation-based mapping and evaluate each technique's efficacy through performance gains in action recognition and decay rate of training loss. We demonstrate that by creating a relevant and more concise knowledge representation, we improve performance (mAP) of action recognition in sports by 26.6% and achieve faster converging models due to higher representation power.

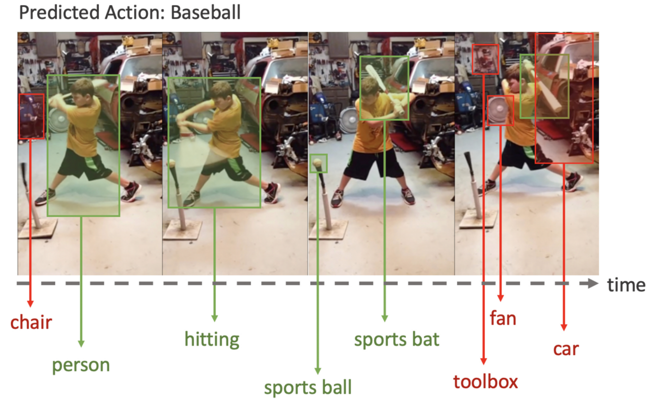

Figure 1. We present scene graph abstraction techniques that rule out irrelevant objects and only encode semantically coherent context that are related to the sport actions. For instance, action "hitting" and object "bat" are semantically relevant to "Baseball" compared with "car" or "chair". Abstracting the relevant information in the feature representation increases the probability of predicting the correct action label. The state-of-the-art video understanding techniques combine long-term symbolic information with short-term visual features [16, 34] for a better representation of the scene. This paper is an extension of their work with the advantage of abstracted scenes.

1. Introduction

The shortcoming of modern perception systems is creating an automated solution that can answer questions such as: Where in the video/image should the classifier pay attention to? Specifically, which objects, subjects, and relations, should be considered in the feature representations? Our approach starts with building scene graphs, as they best describe the scene by localizing the objects, subjects, and their relationships within each frame in a graphical structure. We address one of the main challenges hindering real-world applications using scene-graphs, the problem of generating highly cluttered representations through dense graphs with lots of irrelevant objects and relations. To thoroughly examine the power of relevancy in semantic information and their efficacy in action recognition, we formulate four fundamentally different feature abstraction techniques. In the first technique, we use conditional random fields (CRF) to increase the significance of objects and relations that statistically appear together and reduce those that do not depend on each other (conditioned on each action label). In the second technique, we use canonical correlation analysis (CCA) from a transfer learning perspective to map features to an embedding space where the embedded representations are similar to ground truth action labels. In the third technique, we supply a global semantic context by computing semantic similarity of the action classes to objects and relations and select the most semantically relevant pairs. The fourth technique uses a dimensionality reduction approach based on kernel principal component analysis (kPCA) to select the significant components in feature representation. The state-of-the-art video understanding method aggregates the long-term and short-term visual information through an attention-based mechanism between symbolic-level (like a scene-graphs) and pixel-level visual information [16]. At the core of this paper, we aim to influence their neural network's attention mechanism to give higher weights to the symbolic information that are of significance and semantically relevant (e.g., sport equipment, sport arena) given the hypothesized actions (e.g., throwing, running).In contrast to previous work, our work examines symbolic representation's expressive power through the significance of principal components, semantic similarity, inter-dependencies between objects and relations, and linear mapping of features toward sport actions while referencing the short-term pixel-level information to the long-term symbolic information. Toward this goal, our paper's outline is as follows: in section 2, we go over the related literature. In section 3, we describe the formulation for each abstraction technique. In section 4, we discuss the efficacy of each abstraction technique by feeding the joint representation of abstracted scene graphs and the pixel-level visual appearances to a multi-label action classifier and finally discuss the pros and cons of each technique in section 5.

Contribution Summary: Our main contribution is in examining the efficacy of the four feature abstraction techniques, as each emphasizes a unique attribute within the features:

- CRF-based abstraction, to evaluate significance of object and relations co-occurrence statistics
- CCA-based abstraction, to evaluate the effect of linear projection toward action labels
- Global-based abstraction, to evaluate the overall contextual semantic similarity
- PCA-based abstraction, to evaluate the significant of significant components in the feature representation

2. Related Work

In recent years, the video understanding literature has demonstrated outstanding results due to parallel advances that are converging now. These advances have tried to address two critical points: how to learn the best representation for scene understanding and how to make the model pay attention to the relevant information. In this section, we go over the evolution of video understanding and put our work in perspective.

Convolutional neural networks (CNN) have been powerful models for image representation. Their strength in video understanding has been demonstrated numerously through frame-level to video-level feature extractions [32, 39] with complex spatio-temporal architectures such as 3D ConvNets [30, 31] and effective two stream networks [28, 3] that joint modeling of appearance and motion in videos.

Long-term feature encoding is a relatively new direction [19, 21, 29] and it is complementary to the local CNN features, however, their advantages are not fully exploited partly due to the limitation that a) most datasets contain human activities that span only a few seconds [18, 17, 22] and b) memory constraints. One major difference between these methods is the feature aggregation technique which learns the feature representations through sub-sampling (e.g., 3-7 frames per video) [32, 39].

Scene-graph representations are a symbolic level representation of the scene that describe the object and their relations through graphical structures. There have been a wide range of approaches to generate these graphical representations including [36]: CRF based approaches which model the significance of each object and relations based on co-occurrence statistics [8, 7], Visual transition embedding based approaches which model the translation vector between objects and subjects in different embedding spaces [15, 11] CNN based approaches that exploit the statistical relations over the RoI derived from CNN features [8, 33, 38], Recurrent neural net (RNN) based approaches which form a message passing paradigm and iteratively refine the quality of the scene graphs [35, 6, 37], however, their performance is relatively degraded because they don't consider semantic meaning of objects/relations and and Graph neural net (GNN) based approaches, which aims to factorize the graphs into sub-graphs and refine the scene graph generation through intelligent sub-graph merging. [20, 23, 5]. The scene representation with graphical structure has demonstrated a high descriptive power for many scene understanding tasks. In recent years researchers have evaluated the utility of these symbolic level representations in video understanding by aggregating them with the low-level features extracted through 3D CNNs [16]. Our work is mainly inspired by this study with the extension that our scene graph generation tries to eliminate semantic context that is irrelevant and keep the portion of the features that are significant and discriminative.

Semantic coherence has been a popular topic in natural language processing but relatively underutilized in video understanding. The main drawback is that models built with word embeddings are vulnerable to small perturbations in representation and it may radically alter the semantic meanings of objects. To address this challenge, we encode contextual information such that the semantic compatibilities are conditionally refined based on the scene's global assess-

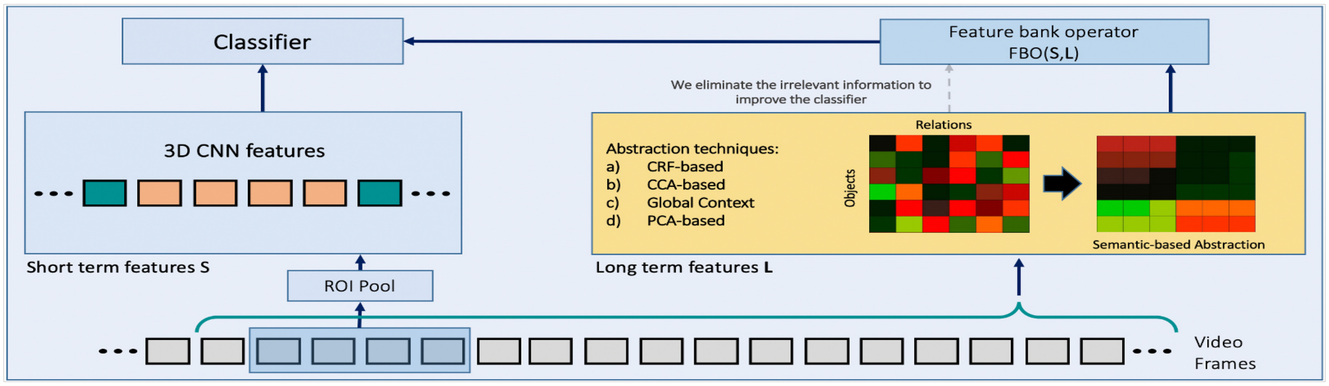

Figure 2. The overall pipeline of aggregating the symbolic-level embedding in long-term features to pixel-level embedding in short term features using semantic coherency. What separates our work from the state of the art [34, 16] is in the abstraction of semantic information which rules out the irrelevant objects and relations before evaluating a scene for classification.

ment. Our work attempts to close this gap using the word embedding concept from [12] by linking the visual context to the abstracted semantic content while aggregating relevant features.

3. Semantic Abstraction

An effective video understanding model should be aware of semantic dependencies between hypothesized actions and the associated temporal context, such as objects and their relations around the time the action is taking place. In this section, we discuss four fundamentally different techniques for capturing such dependencies between the actions and the objects/relations associated to each person. We investigate these dependencies based on: a) co-occurrence statistics between actions and object relations using conditional random fields (CRF), b) based on a linear projection of the object/relation features using canonical correlation analysis (CCA), c) based on principal components analysis (PCA) of features that are computed over objects and relations, d) and based on semantic similarity to action labels (global semantic context).

At a high level, our goal is to provide the model with features relevant to the hypothesized actions and to assess the relevance of these abstracted features by monitoring the performance gain as the model searches for important cues and eliminates irrelevant context from the feature representation of the scene. In this paper, we set up our baseline classification model similar to [34] as they have already demonstrated that grounding the objects and relations in a structured feature representation can improve the performance of the action classifier. We now go over the feature bank generation with scene graphs followed by the problem statement before describing each abstraction technique.

Scene-graph feature bank generation is the initial stage of the process, where for each video frame f_i, a graphical representation $G = (O, R)$ of all objects, $O = o_1, o_2, o_3, ..., o_m$, and all relations, $R = r_{11}, r_{12}, ..., r_{21}, r_{22}, ...$, is captured between all detected actors in the scene such that r_{pq} is the relation between the p^{th} object o_p and the q^{th} object o_q. We select only the relations for which class of either the subject or the object in the subject-relation-object triplet is a person. Furthermore, we obtain the associated confidence probabilities for the prediction of each objects $s_1, s_2, ...$ and relations $s_{11}, s_{12}, ..., s_{21}, s_{22}, ...$ Given the confidences scores of the objects and relations associated with each person we construct a confidence map $C_{ij} = s_i \times s_{ij}$ which captures the confidence for each combination of object relation for a given actor in the scene. We then flatten each matrix C to generate each element of the scene graph feature bank $F_{SG} = [f_1, f_2, ..., f_T]$, where T is the total number of time steps while f_t is the flattened C matrix at time t.

Problem Statement

Given an entire video and a set of action labels L, the objective is to assign the label $l_t \in L$ to each video frame, (note that the classification could potentially be performed on a small duration of the videos as well). Our classification framework is similar to [34, 16], which trains an attention-based neural network architecture that references the short-term information to the long-term information using 3D CNN features and the scene graph feature banks F_{SG}. Given this framework, our goal is to modify the feature banks to adjust relevant objects $\hat{o}$ and relations $\hat{r}$ such that,

$$F'_{SG} \vdash \underset{l}{\text{argmax}}\ P(l_t, \hat{o}, \hat{r} | F_{SG}, t) \tag{1}$$

Where F'_{SG} is the abstracted feature with encoded relevant object/relations. Next, we discuss different techniques used for the abstraction of the raw feature banks F_{SG} to F'_{SG} to eliminate irrelevant objects and relations before feeding the features to the action classifier. We then discuss the efficacy of each technique on the performance of action recognition.

3.1. CRF-based Abstraction

We now formulate a conditional random field (CRF) on top of the feature bank representation F_{SG} with object set O, where $\{o_i \in O\}_{i=1}^{N}$ and relation set R, where $\{r_j \in R\}_{j=1}^{M}$. A typical CRF formulation involves constructing a graphical model $G = (O, R)$, where O is a set of N vertices while R is a set of $\binom{N}{2}$ relationship edges. For simplicity, only the unary and the second-order interactions are considered. The conditional probability distribution of label l given the input feature F_{SG} can then be written as

$$P(l, \hat{o}, \hat{r} | F_{SG}, v_t) = \frac{\exp\{\sum_{i=1}^{N} \Psi_u(o_i, \boldsymbol{\theta}) + \sum_{i=1}^{N} \sum_{j=1}^{M} \Psi_p(r_{ij}, \boldsymbol{\omega})\}}{Z(\mathbf{o}, \mathbf{r})} \tag{2}$$

where, $\boldsymbol{\theta} = [\theta_i, ..., \theta_N]$, and $\boldsymbol{\omega} = \begin{bmatrix} \omega_{00} & & \\ & ... & \\ & & \omega_{NM} \end{bmatrix}$

$\Psi_u(\cdot)$ measures the unary cost of selecting a particular object and $\Psi_p(\cdot)$ measures the pairwise cost of selecting particular relation conditioned on the action labels. θ_i denotes the weight associated with the i^{th} object and ω_{ij} denotes the weight associated with i^{th} object and j^{th} relation. The edge parameters $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$ are obtained from co-occurrence statistics while predicting the labels. Thus, the mutual information is estimated implicitly through most common objects and relations for each action label. Given each instance of the video segment and the corresponding feature bank F_{SG}, the *unary potentials* $\Psi_u(\cdot)$ capture the discriminative power of each node (objects) and *pairwise potentials* $\Psi_p(\cdot)$ capture the discriminative power of each edges (relations). Each edge between i and j is characterized by M^2 connections, each representing the relevance to particular object-relation combination. For instance, the likelihood of the label l given $o_i = 0$ and $r_{ij} = 1$ is captured by the weight θ_0 and ω_{01}. Consequently, in order to map the feature banks to action labels in a given frame, $N + \binom{M}{2} \times N^2$ potentials are aggregated over the entire objects and relations to make a prediction. Sum of all potentials form an un-normalized distribution, and the normalizing partition function $Z(\cdot)$ is used to form a probability distribution over the sum.

Unary and Pairwise Potentials: In order to capture the relevance of each object in the scene we compute the probability of each objects and relation pairs given the labels. With scene-graph model, we directly use the confidences of predicted labels s_i as a proxy measure to assess the relevance of each object given the labels, hence the unary potentials are defined such that:

$$\Psi_u(o_i, \boldsymbol{\theta}) = -\log\{\theta_i f_i(o_i)\} \tag{3}$$

where the node feature functions are defined as $f_i(o_i) = P_n(\hat{o}, \hat{r} | F_{SG}, v_t) = s_{o_i}$. $P_n(\cdot)$ is the confidence measure associated with the object o_i and θ_i reflects the significance of each object which is learned based on co-occurrence statistics for a given dataset. Referring back to the conditional likelihood model (Eq. 2), to make the prediction, each pairwise potential $\Psi_p(\cdot)$ is summed over the entire graph G and the likelihood of a label given each possible edge is formulated as a negative log of logistic regression classifier. Therefore the edge feature functions are defined as $f_{ij}(\cdot)$ as the probability of each specific relation given the labels. i.e.,

$$\Psi_p(r_{ij}, \boldsymbol{\omega}) = -\log\{\frac{1}{1 + \exp\{\beta_{ij}^{r_{ij}} + \omega_{ij}(r_{ij}) f_{ij}^{e}(r_{ij})\}}\} \tag{4}$$

$$f_{ij}^{e}(r_{ij}) = -\log P_2(l \mid r_{ij})$$

$P_2(\cdot)$ are the co-occurrence statistics of a given the relationship pairs for a particular label while β is the bias. Once the parameters are estimated for each label we use mAP inference for each test sample and pick the label in which its parameters returns the highest conditional likelihood, thus we rewrite Eq. 1 such that :

$$\hat{l} = \underset{l}{\operatorname{argmax}} \exp\{\sum_{i=1}^{N} \Psi_u(o_i, \boldsymbol{\theta}) + \sum_{i=1}^{N} \sum_{j=1}^{M} \Psi_p(r_{ij}, \boldsymbol{\omega})\}$$

Inference/parameter estimation: The parameter estimation technique for such CRF model depends highly on the complexity of the structure. The goal here is to estimate a set of weights $(\boldsymbol{\theta}, \boldsymbol{\omega})$ that maximizes the accuracy of our prediction given the labels. If the number of parameters are below a certain threshold we use the stochastic gradient ascent method to maximize $P(l, \hat{o}, \hat{r} | F_{SG}, v_t, \boldsymbol{\theta}, \boldsymbol{\omega})$. In this case the weights are updated using standard gradient ascent. For more complex structures, CRF distribution is approximated with mean field approximation where iterative message passing is performed for approximate inference.

3.2. CCA-based Abstraction

The aforementioned C_{ij} matrices are two-dimensional confidence maps of objects and relations interacting with each person in the scene. These confidence maps lay on a linear manifold, in the sense that the linear (or convex) combination of two confidence maps could reasonably belong to the set of confidence maps obtained from a similar dataset. Motivated by this, our goal is to find a linear mapping from the observed semantic content (represented as C_{ij}) to the multi-hot vector obtained from ground truth action labels. After flattening the confidence maps to F_{SG}, the relative information is embedded into two vectors, namely $\mathbf{h}$ and $\mathbf{v}$,

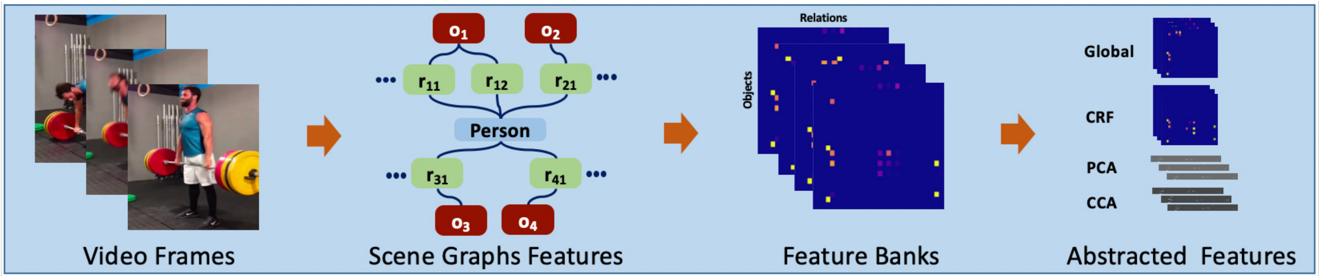

Figure 3. Feature bank abstraction process for a video from SVW dataset. The semantic label for this clip depicts "weight lifting." First, the scene graph representation is obtained which contains all objects and relations. Next, we construct a raw feature bank matrix where the first axis holds all objects that the actor is interacting with and second axis holds the actor's relation when interacting with the these objects. The raw feature banks are then fed to four independent abstraction process generating a unique representation containing different attributes. These features are then used to augment the 3D CNN features as input to the classifier.

(for simplicity, in this abstraction technique the F_{SG} vectors are denoted as $\mathbf{h}$ and the multi-hot vectors are denoted as $\mathbf{v}$). Let N be the total number of features used for training such that $\mathbf{h_t} \in \mathbb{R}^{(\mathbf{M*N})\times\mathbf{1}}$ and $\mathbf{v_t} \in \mathbb{R}^{|\mathbf{L}|}$, where $M * N$ is the length of the first dimension of the F_{SG} and $|L|$ is the size of action labels used to create the multi-hot vectors. We want to find the relationship between the observed scene graph representation and the multi-hot vectors from ground truth labels by finding a lower-dimensional subspace in which the $\mathbf{v}$ and $\mathbf{h}$ are most correlated. In other words, the projection of $\mathbf{u}^T\mathbf{h}_t$ and the corresponding multi-hot vectors $\mathbf{v}_t^T\mathbf{w}$ into the shared subspace are highly correlated. For this purpose, we use Canonical Correlation Analysis (CCA) for such mapping. CCA seeks a shared embedding for $\mathbf{h}$ and $\mathbf{v}$ such that the embedded representations for the same instances lie close to each other and subsequently maximizes the following objective function:

$$
\begin{aligned}
&CCA_{comp} = \\
&\underset{\mathbf{u},\mathbf{w}}{\operatorname{argmax}} \frac{\sum_{n=1}^{N}(\mathbf{u}^T h_t)(\mathbf{v}_t^T\mathbf{w})}{\sqrt{\sum_{n=1}^{N}\mathbf{u}^T\mathbf{h}_t\mathbf{h}_t^T\mathbf{u}}\sqrt{\sum_{n=1}^{N}\mathbf{w}^T\mathbf{v}_t\mathbf{v}_t^T\mathbf{w}}} \\
&\quad = \underset{\mathbf{u},\mathbf{w}}{\operatorname{argmax}} \frac{\mathbf{u}^T C_{hv}\mathbf{w}}{\sqrt{\mathbf{u}^T C_{hh}\mathbf{u}}\sqrt{\mathbf{w}^T C_{vv}\mathbf{w}}}
\end{aligned}
$$

where $\mathbf{u}$ and $\mathbf{w}$ are the CCA components which project the data onto the shared embedding and C_{hh}, C_{vv}, C_{hv} are the variance matrices.

3.3. Global Context based Abstraction

While the contextual information of the scene has been modeled implicitly through the bottom-up local operators of CNNs, the explicit aggregation of relevant semantics is not well captured. The utility of semantic coherence has been evaluated in many tasks before [12][24]. Similarly, we use the notion of semantic similarity but to abstract the feature banks F_{SG} such that objects and relations that are not semantically relevant get eliminated and a smaller feature bank $F'_{SG} \in \mathbb{R}^{k\times L}$ is obtained with top k significant object-relation pairs.

We use [24] to encode sentences sen_{ij} and sen_l to obtain sentence embeddings e_{ij} and e_l. Here, sen_{ij} is the sentence representing a textual description of object-relation pair, $\langle o_i, r_{ij}\rangle$, with object o_i and relation r_{ij} while sen_l is the textual description of the action label l. For instance, if object o_i is a "television" and relation r_{ij} is "watching a', then sen_{ij} is "watching a television". For actions, sen_l is the textual description of the action label like "Sitting on a table" or "Sitting in a chair". For each action l, we sort all object-relation pairs using, as a key, the cosine similarity of the action sentence sen_l with each object-relation sentence sen_{ij}. We take the top k object-relation pairs for each action label, l to obtain kL object-relation pairs,

$$
\begin{bmatrix}
\langle o,r\rangle_{0,0} & \langle o,r\rangle_{0,1} & \dots & \langle o,r\rangle_{0,l} \\
\langle o,r\rangle_{1,0} & \langle o,r\rangle_{1,1} & \dots & \langle o,r\rangle_{1,l} \\
\vdots & \vdots & \ddots & \\
\langle o,r\rangle_{k,0} & \langle o,r\rangle_{k,1} & \dots & \langle o,r\rangle_{k,l}
\end{bmatrix}
$$

that are used for global semantic context based abstraction. For every frame, we prune F_{SG} by discarding all object-relation pairs that are not among the aforementioned kL global pairs to obtain abstracted F'_{SG} with global semantic context.

3.4. Kernel-PCA based Abstraction

A good knowledge representation can lead to a faster and more accurate inference model [2]. Motivated by these, we consider experimenting with PCA through dimensionality reduction. More specifically, we apply kernel PCA (kPCA) [13] which is a nonlinear extension of PCA that has the capability to exploit redundancies through higher order statistics, for a relevant object/relation abstraction. The principal components of the PCA are computed with the eigen-decomposition of the covariance matrix $\Gamma = \frac{1}{n}F_{SG}^T F_{SG}$,

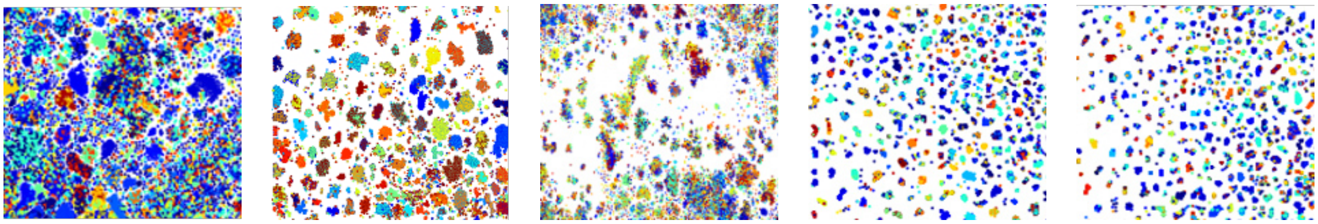

Figure 4. Sampled tSNE plots from each feature bank representations: from left to right, a) raw feature banks, b) CCA-based features, c) CRF-based features, d) PCA-based features and e) global-context based features. Notice that the separability between clusters are more distinct in abstracted representations compared with the basic feature representation.

which decomposes the covariance matrix to $\Gamma = Q\Lambda Q^T$ as Q depicts the direction of maximum variance. Such decomposition for the kernel-base version happens in the feature space (rather than input space [10]), instead, we have $\hat{\Gamma} = \frac{1}{n}\Phi^T\Phi\mathbf{q} = \lambda\mathbf{q}$ which leads to a decomposition of $\frac{1}{n}\Phi^T\Phi = Q\Sigma Q^T$. With the span of $\{\Phi^T\}$ and $\{Q\}$ being equal, each vector q can be written in terms of n-dimensional vector $\mathbf{p}$ as $\mathbf{q} = \mathbf{\Phi^T p}$. Which then by assuming $K\mathbf{p} = \lambda\mathbf{p}$, we have the decomposition of $K = P^T$. Depending on the complexity of the feature banks and their dimension, there are different alternatives to estimate the eigenpairs (both exact and approximation techniques). In this work we use an approximation method that works based on iterative improvements toward an exact solution.

4. Experiments

In this section, we evaluate the efficacy of each abstraction technique in the action recognition framework. We train and test our models on four Nvidia Quadro RTX 8000 GPUS. We use two datasets to evaluate our methods: Charades and Sports Videos in the Wild (SVW). SVW, as compared with Charades, does not have as many objects and relations present in the scene. Therefore, we first evaluated the efficacy of our methods in Charades and then picked the best feature representation (S_P) and abstraction technique (PCA-based) to evaluate action recognition performance.

4.1. Semantically Weighted Feature Banks

For each video segment we first localize all persons, objects and relations using [9]. We then compute the confidence map of all objects and relations for every frame (sampling takes takes place during training and testing of the classifier). Next we enhance the representation power of F_{SG} with four different representations. The first is the original scene graph feature bank C. In C_p, we multiply the scene graph matrix C by the detection confidence of the actors. In the third representation, S, we scale each element of C by the semantic distance between each object-relation pair computed using the ViCO word embeddings [12]. In S_p, we scale S by the confidence of each actor. In our experiments we evaluated all four scene graph matrix types C, C_p, S, S_p (Table) and used the most discriminate representation to evaluate our models.

Feature Bank Variations	
C	71.9
C_p	72.2
S	71.8
S_p	72.2

Evaluation on Charades Dataset	
PCA	72.2
CCA	72.2
Global-context	72.0
CRF	71.8
STO [34]	71.7

Table 1. Performance on different abstractions. Left table shows the effect of using different scene graph matrices (comparisons are shown for the best performing abstraction technique). C is from the original scene graph object-relation matrices, C_p is C scaled with detection confidence of actors, S is adjusted C based on semantic distance between each object-relations, and S_p is S scaled with confidence of actor detection. We compare the performance (mAP) across top 30 action labels using S_p. See 4.1 for the definitions of the S_p matrix.

CRF setup: We formulate a conditional random field such that the unary potentials are represented with the most common objects, and the pairwise potentials are represented with the most common relationship pairs conditioned on the action labels. The feature function of unary potentials uses the confidence of object detection directly, and the feature function of pairwise potentials uses both object and relationship confidences to produce discriminative features. In situations where the data structure is too large and highly sparse, we can alternatively use the limited memory quasi-Newton technique for bound-constrained optimization to better estimate unary and pairwise weights.

Deep-CCA setup: We first convert the ground-truth labels into multi-hot vectors, and apply a non-linear extension of CCA [1] which transforms the feature banks and multi-hot vectors to a new embedding space where their correlations are maximized. Because the multi-hot vectors already have the most discriminative form with respect to the ground-truth labels, our procedure maps the features banks to be more discriminative.

PCA setup: For this technique, the goal is to eliminate the least significant components of the feature banks. Therefore, we apply kernel PCA to F_{SG}, which has a size of 7701, and select the first 2048 PCA components to represent the scene. Note that unlike CRF or Global-context based techniques, the explicit notions of object and relations are lost in this abstraction (as shown in Fig. 3), but by eliminating the least significant components of the feature bank we aim to eliminate noise in our feature banks.

Global-context setup: We compute the semantic similarity between each action and each object-relation pair in F_{SG} as described in 3.3. The globally abstracted feature bank F'_{SG} is then obtained by using the top k values in the semantic similarity matrix. The abstracted feature from this method consists of the top k similar object-relation pairs to each action class. The globally abstracted feature bank F'_{SG} is then obtained by using only the top k similar object-relation pairs to each action class. In our experiments, we empirically found the best k to be equal to 13.

4.2. Action Recognition with Abstracted Features

We split each video into clips and pass each clip through I3D [4] with a ResNet101 backbone [14] to compute a short-term feature of dimension 2048. Instead of computing a long-term feature bank (LFB), we used our abstracted features banks mentioned above. For each clip, we generate all neighboring F'_{SG} in a window of size $2d+1$ centered at the current clip. We then use the same Non-Local Feature Bank Operator (FBO-NL) as [34] to aggregate our abstracted features F'_{SG} with the short-term features before passing them through the classifier (as illustrated in Fig.2).

SVW: The SVW (Sport Videos in the Wild) dataset [25] consists of 4,200 videos captured by a smartphone. There are 30 sport categories and 44 different actions. Each video has a single sport label and 40% of the video is labeled in time and space with a single action. We split the data into 75% train, 25% validation set, and sample the SGFB every 10 frames. Our mAP across 30 sport actions is 88.1% which was a significant improvement compared to reported performance of 61.5% in [26].

Charades Dataset[27]: Charades dataset consists of 9,848 videos, which are on average 30 seconds. There are 157 action classes and each video can be labeled with multiple actions. We use 7811 videos for training and 1814 for testing and the remainder are pruned. We extract 32 frame clips from each training/testing video with a stride of four frames. We use a window size of $d = 10$ for our feature bank. We use the same 3D CNN backbone, hyperparameters, and optimizer as previous work [34, 16] for a fair comparison.

4.3. Comparing Abstractions Techniques

We noticed that the classification performance of action labels depends on different abstraction technique. Some actions, specially in sports, involve specific equipment. Our abstraction techniques show better performance for such actions. In CRF and global-context based technique we explicitly encode relevant objects/relations, compared with

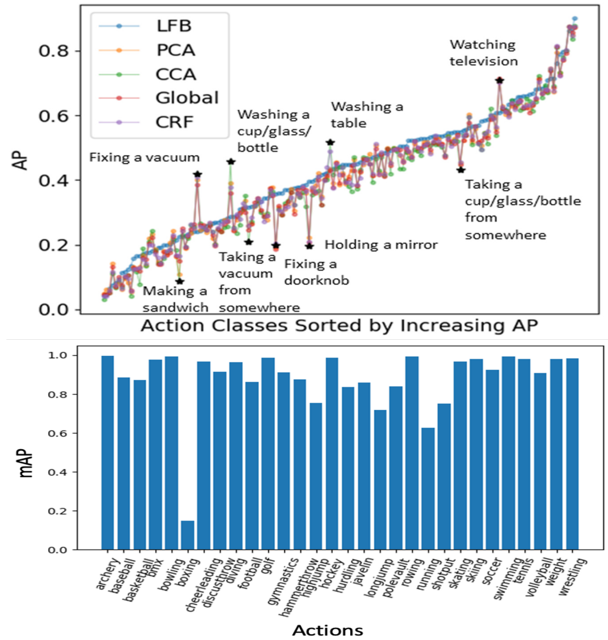

Figure 5. Top: Action recognition performance on Charades. The performance variation between different abstraction is due to whether actions involve relative equipment and how common those equipment are between actions. Bottom: Performance of best model (PCA) for action recognition on SVW

CCA based techniques where the relevancy is implicit in the linear mapping. Using CRF, Global and CCA based method we are able to explicitly remove objects and relations that are not related to the actions. This makes it suitable for datasets where people are interacting with equipment. In PCA based technique we abstract the features using the most significant component of the feature. In comparison to other techniques, the PCA technique is advantages when there aren't many object and relations present in the scene.

5. Conclusion

We examined four fundamentally different feature abstraction techniques to improve action recognition for sports. We were able to significantly improve the action recognition mAP on SVW by 26.6% through automatic abstraction of relevant information in the scene. In summary, each abstraction technique is based on a unique criterion and has different effects on action recognition performance. With the global-context and the CRF technique we explicitly exploit the object relations whereas with PCA and CCA technique we implicitly obtain significant components or correlated mapped features. Through our experiments, we were able to express the feature representations much more efficiently, consequently leading to more accurate and faster converging classification models.

References

[1] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR, 2013. 6

[2] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019. 5

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conf. on CVPR*, pages 6299–6308, 2017. 2

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conf. on CVPR*, 2017. 7

[5] Long Chen, Hanwang Zhang, Jun Xiao, Xiangnan He, Shiliang Pu, and Shih-Fu Chang. Counterfactual critic multi-agent training for scene graph generation. In *IEEE Int. Conf. on CV*, pages 4613–4623, 2019. 2

[6] Yunian Chen, Yanjie Wang, Yang Zhang, and Yanwen Guo. Panet: A context based predicate association network for scene graph generation. In *2019 IEEE Int. Conf. on Multimedia and Expo (ICME)*, pages 508–513. IEEE, 2019. 2

[7] Weilin Cong, William Wang, and Wang-Chien Lee. Scene graph generation via conditional random fields. *arXiv preprint arXiv:1811.08075*, 2018. 2

[8] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *IEEE Conf. on CVPR*, pages 3076–3086, 2017. 2

[9] Kaihua Tang et al. Unbiased scene graph generation from biased training, 2020. 6

[10] Schölkopf et al. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998. 6

[11] et al. Gkanatsios. Deeply supervised multimodal attentional translation embeddings for visual relationship detection. In *2019 IEEE Int. Conf. on Image Processing (ICIP)*, pages 1840–1844. IEEE, 2019. 2

[12] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. Vico: Word embeddings from visual co-occurrences. In *IEEE Int. Conf. on CV*, pages 7425–7434, 2019. 3, 5, 6

[13] Fredrik Hallgren and Paul Northrop. Incremental kernel pca and the nystr\" om method. *arXiv preprint arXiv:1802.00043*, 2018. 5

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. on CVPR*, 2016. 7

[15] Zih-Siou et al. Hung. Contextual translation embedding for visual relationship detection and scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[16] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *IEEE Conf. on CVPR*, pages 10236–10247, 2020. 1, 2, 3, 7

[17] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conf. on CVPR*, pages 1725–1732, 2014. 2

[18] et al. Kay. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2

[19] et al. Li, Fu. Temporal modeling approaches for large-scale youtube-8m video understanding. *arXiv preprint arXiv:1707.04555*, 2017. 2

[20] Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. Factorizable net: an efficient subgraph-based framework for scene graph generation. In *European Conf. on CV (ECCV)*, pages 335–351, 2018. 2

[21] Antoine Miech, Ivan Laptev, and Josef Sivic. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905*, 2017. 2

[22] et al. Monfort. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019. 2

[23] Mengshi Qi, Weijian Li, Zhengyuan Yang, Yunhong Wang, and Jiebo Luo. Attentive relational networks for mapping images to scene graphs. In *IEEE Conf. on CVPR*, pages 3957–3966, 2019. 2

[24] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conf. on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. 5

[25] Seyed Morteza Safdarnejad, Xiaoming Liu, Lalita Udpa, Brooks Andrus, John Wood, and Dean Craven. Sports videos in the wild (svw): A video dataset for sports analysis. In *Int. Conf. on Automatic Face and Gesture Recognition*, 2015. 7

[26] Seyed Morteza Safdarnejad, Xiaoming Liu, Lalita Udpa, Brooks Andrus, John Wood, and Dean Craven. Sports videos in the wild (svw): A video dataset for sports analysis. In *Proc. International Conference on Automatic Face and Gesture Recognition*, Ljubljana, Slovenia, May 2015. 7

[27] Gunnar A. et al. Sigurdsson. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conf. on CV*, 2014. 7

[28] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014. 2

[29] Yongyi Tang, Xing Zhang, Lin Ma, Jingwen Wang, Shaoxiang Chen, and Yu-Gang Jiang. Non-local netvlad encoding for video classification. In *European Conf. on CV (ECCV)*, pages 0–0, 2018. 2

[30] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *IEEE Int. Conf. on CV*, pages 4489–4497, 2015. 2

[31] Gül et al. Varol. Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1510–1517, 2017. 2

[32] et al. Wang. Temporal segment networks: Towards good practices for deep action recognition. In *European Conf. on CV*, pages 20–36. Springer, 2016. 2

[33] Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon. Linknet: Relational embedding for scene graph. In *Advances in Neural Information Processing Systems*, pages 560–570, 2018. 2

[34] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *IEEE Conf. on CVPR*, pages 284–293, 2019. 1, 3, 6, 7

[35] et al. Xu. Scene graph generation by iterative message passing. In *IEEE Conf. on CVPR*, pages 5410–5419, 2017. 2

[36] Pengfei Xu, Xiaojun Chang, Ling Guo, Po-Yao Huang, Xiaojiang Chen, and Alexander G Hauptmann. A survey of scene graph: Generation and application. Technical report, EasyChair, 2020. 2

[37] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *IEEE Conf. on CVPR*, pages 5831–5840, 2018. 2

[38] Ji Zhang, Mohamed Elhoseiny, Scott Cohen, Walter Chang, and Ahmed Elgammal. Relationship proposal networks. In *IEEE Conf. on CVPR*, pages 5678–5686, 2017. 2

[39] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *European Conf. on CV (ECCV)*, pages 803–818, 2018. 2