# Modeling Scene and Object Contexts for Human Action Retrieval with Few Examples

Yu-Gang Jiang, Zhenguo Li, and Shih-Fu Chang, *Fellow, IEEE*

***Abstract*—The use of context knowledge is critical for understanding human actions, which typically occur under particular scene settings with certain object interactions. For instance, *driving car* usually happens outdoors, and *kissing* involves two people moving toward each other. In this paper, we investigate the problem of context modeling for human action retrieval. We first identify ten simple object-level action atoms relevant to many human actions, e.g., *people getting closer*. With the action atoms and several background scene classes, we show that action retrieval can be improved through modeling action-scene-object dependency. An algorithm inspired by the popular semi-supervised learning paradigm is introduced for this purpose. One important contribution of this paper is to show that modeling the dependencies among actions, objects, and scenes can be efficiently achieved with very few examples. Such a solution has tremendous potential in practice as it is often expensive to acquire large sets of training data. Experiments were performed on the challenging Hollywood2 dataset containing 89 movies. The results validate the effectiveness of our approach, achieving a mean average precision of 26% with just ten examples per action.**

***Index Terms*—Action retrieval, context modeling, object and scene recognition, very few examples.**

## I. Introduction

Human ACTION recognition is one of the most challenging problems in video analysis due to significant appearance variations of human bodies, background clutter, and occlusion. Although impressive results have been reported from evaluations on datasets collected from controlled environments, such as KTH [1] and Weizmann [2], much less efforts have been made to solve the more difficult problem of recognizing actions in realistic videos from unconstrained environments [3]–[5]. Furthermore, most existing methods require large sets of labeled training samples, which are difficult and expensive to acquire in practice.

In this paper, we are interested in the problem of finding realistic human actions based on a limited number of examples, with a focus on modeling contextual cues associated with the actions. Human actions are defined by their appearances with certain motion characteristics, and generally occur under particular scene settings. For example, as shown in Fig. 1, action *kissing* contains two people with a shrinking distance between them. The *action-scene-object dependency* provides rich contextual information for understanding human actions. Most previous approaches, however, handled actions, scenes, and objects separately without considering their relationship [3], [6]. Our intuition is that once the contextual cues (objects and scenes) are learned *a priori*, they can be utilized to make the action retrieval more effective. We, therefore, explore such contexts in this paper, aiming at modeling arbitrary realistic human actions rapidly from just a few examples, instead of relying on sophisticated models learned from hundreds or even thousands of labeled samples for only a limited set of predefined action categories.

We first identify a set of object-level action atoms that can be automatically detected based on the recent progress in object detection [Fig. 1(a)]. These object-level atoms include both single object and multiple objects with certain interactions. With the action atoms and a few background scene classes that are relevant to many human actions, we introduce an efficient algorithm to capture the underlying action-object-scene dependency. The algorithm, which couples the advantages of both maximum margin learning and semi-supervised learning, allows us to utilize the large amount of readily available unlabeled data. The contextual dependency information is then harnessed for finding similar actions from new video data.

Our main contribution in this paper is to show that, by tailoring popular ideas from machine learning, action-scene-object dependency can be modeled from very few examples, and the contextual cues from both object detection and scene categorization could be effectively harnessed for human action retrieval. To the best of our knowledge, this is the first work on efficient modeling of action context using limited examples. Our evaluation on the challenging Hollywood2 dataset shows that our method accurately models the action context, producing competitive performance compared to the state of the art using a domain-specific context prediction method based on textual mining in movie scripts.

## II. Related Work

Human action recognition has been an important research topic in video analysis and computer vision with a vast amount of applications, such as video search, robotics, and surveillance. Focuses of existing works ranged from feature

Manuscript received August 1, 2010; revised November 9, 2010; accepted December 26, 2010. Date of publication March 17, 2011; date of current version May 4, 2011. This work was supported by the NGA Award HM1582-09-1-0036, the NSF Award CNS-07-51078, and a research gift from Kodak. This paper was recommended by Associate Editor E. Izquierdo.

The authors are with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA (e-mail: yjiang@ee.columbia.edu; zgli@ee.columbia.edu; sfchang@ee.columbia.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2011.2129870

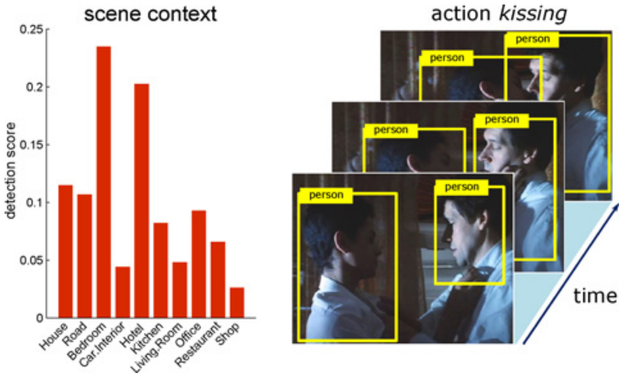


Fig. 1. Proposed approach models both object interactions and background scenes for retrieving human actions such as *kissing* shown in the above video clip. We first detect objects such as (a) *Person* on the video sequence using state-of-the-art object detectors and (b) classify video scenes using several pre-trained models. With a few labeled action video clips, our approach is able to identify the contextual dependencies of the action *kissing* to object interactions such as *people getting closer*, and scenes such as *Bedroom*. The contextual dependency information is then utilized for finding *kissing* in new videos.

representation [3], [4], [7]–[11] to recognition model [1], [12], [13]. All the works in [7]–[9] are on the design of spatial-temporal features, while more recent research in [3], [4], [10], and [11] focuses on higher level event representation. In this paper, we use the popular bag-of-features framework which describes actions by histograms of quantized local features [3], [8]. We adopt local features extracted from both 2-D video frames [scale-invariant feature transform (SIFT) [14]] and 3-D sequences (spatial-temporal interest points [7]).

Extensive research has been devoted to modeling context for visual recognition. Torralba *et al.* [15] introduced a framework for context modeling based on the correlation between overall statistics of low-level image features and objects in the images. Several recent research works along this line adopted context for understanding objects and scenes in static images, e.g., [16] and [17], among others. In [16], the co-occurrence information of objects and scenes was utilized by Robinovich *et al.* for object recognition using a conditional random field framework. Similarly, Russell *et al.* [17] used scene matching and incorporated object spatial priors in object recognition.

Context has also been explored in several recent works for human action recognition [5], [10], [18]–[22]. Gupta *et al.* [18] employed a Bayesian network to model human–object interactions for action understanding. Wu *et al.* [19] investigated contexts from radio-frequency identification-tagged objects for kitchen activity recognition. Tran *et al.* [20] used Markov logic networks (MLN) to impose common sense rules for action recognition in a parking lot. MLN was also applied in [13] for recognizing objects based on human activity context in home environment. Most of these papers [13], [18]–[20], however, only considered video data collected from controlled or very specific environments. Our paper is more similar in spirit to [5], [10], [21], and [22] where contexts were exploited for recognizing actions from realistic video data. Sun *et al.* [10] modeled action context in a feature representation that is designed to capture the proximity of local keypoint tracks. In [5], Marszałek *et al.* exploited scene context for action recognition in movies. Action-scene relations were automatically discovered from video-aligned textual scripts. In [21], context from generic object detectors was applied for realistic action recognition. Responses of the object detectors are quantized into low-dimensional video descriptors which are used as input to supervised action learning. In [22], object and human pose contexts are modeled for detecting activities containing human–object interactions. Unlike [5], [10], [21], and [22], however, we model both scene and object contexts for action retrieval with limited examples, instead of relying on large sets of fully labeled training data. An algorithm derived from semi-supervised learning is introduced for this purpose to model the action context.

Recognizing visual categories from a limited amount of data has been investigated in [11] and [23]. Fei-Fei *et al.* [23] proposed a Bayesian formulation for learning object categories from very few images. In [11], Seo *et al.* proposed to use features based on space-time locally adaptive regression kernels for detecting human actions based on a single example. Our approach is related to [23] in the sense that we also try to utilize models learned previously. However, in our case, both background scene and object-interaction contexts, not pre-trained object category models, are explored for retrieving human actions.

## III. Modeling Actions with Context

Our goal is to model multiple contextual cues for action retrieval from few examples. Fig. 2 overviews the entire framework, which contains four major components. Given a few action examples, the first component automatically mines a number of negative examples. Notice that this step is important under the retrieval scenario where only a few positive samples are given. After that the second component extracts scene-level and object-level contextual cues using existing models/classifiers. With the contextual cues from the positive/negative examples, the third component models the action-scene-object relationship, which is finally utilized in the last component to discover similar actions from new data. In the following, we describe each of the components in detail.

### *A. Video Representation and Negative Sample Selection*

We now briefly introduce video feature representations and the method for automatically collecting negative samples. We follow several existing works [3], [5], [21] to use the bag-of-features framework. Local features are extracted from both 2-D video frames and 3-D video sequence. For the 2-D features, we use Harris-Laplace detector [24] and SIFT descriptor [14]. The 3-D spatial-temporal interest points are extracted based on the work of Laptev *et al.* [3], [7]. Histograms of gradient (HoG) and histograms of optical flow (HoF) are applied to describe the 3-D interest points [3], and the descriptors by HoG and HoF are concatenated into a final descriptor (HoG–HoF; 144 dimensions). For both SIFT and HoG–HoF, we sample a set of 600 000 descriptors, and use $k$-means clustering to generate a codebook of 4000 visual words. Visual codebooks of similar size have been shown to produce good results for a wide range of human action datasets in recent works [3], [5], [25]. With

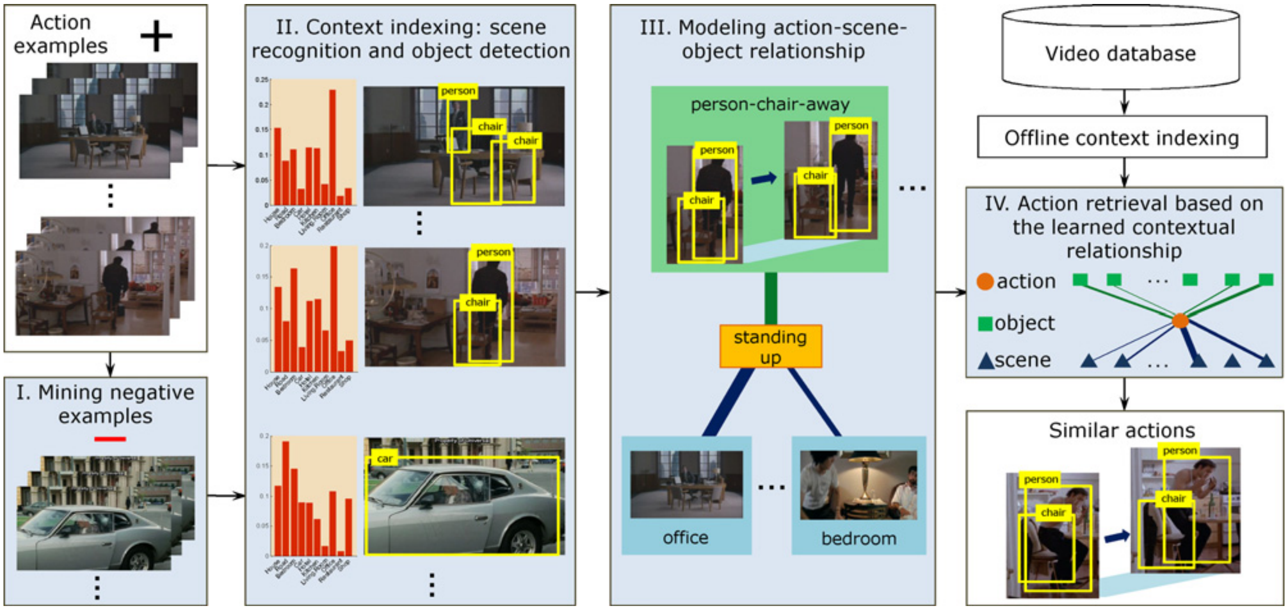


Fig. 2. Overview of our context-based action retrieval framework. Given a few example video clips of an action (*standing up*), we first automatically mine a number of pseudo-negative samples (Section III-A). We then extract object and scene contexts (Section III-B) and predict their relationships to the action (Section III-C). The learned action-scene-object relationship is finally utilized to incorporate multiple contextual cues for discovering similar actions (Section III-D).

the codebooks, each video clip can be represented by two 4000-D histograms of visual words through soft-quantization [26] of the local descriptors. Since the videos are of different length, the visual word histograms are $l_1$-normalized.

As the aim is to retrieve actions from limited examples, our framework only requests positive action examples and automatically collects negative samples from unlabeled video data. To this end, we apply a popular semi-supervised learning algorithm called local and global consistency (LGC) [27]. LGC propagates the given positive examples to the unlabeled data based on manifold structure of all the samples in feature space (HoG–HoF is adopted). The unlabeled samples can then be ranked according to their scores after the propagation, and we randomly pick pseudo-negative samples that are far away from the positive ones. One main advantage of semi-supervised learning is that it utilizes the large amount of readily available unlabeled data. Semi-supervised learning has shown excellent performances in many applications especially when the number of examples (labels) is small, which perfectly fits the aim of this paper. Notice that we avoid using the movie script mining method in [3] to collect labeled samples, as it is particularly designed for movies and may not be applied for other types of action videos. While in the experiments the Hollywood2 dataset is adopted for the ease of performance evaluation, we do not want to limit our approach to any particular type of data.

### B. Obtaining Action Context

1) *Scene Recognition:* Background scene setting is an important source of context for understanding human actions. We adopt ten scene classes defined in [5] [see Fig. 1(b) for class names], and the scene models are learned using one-against-all strategy by support vector machines (SVM). We use the popular $\chi^2$ Gaussian kernel. The effectiveness of this kernel has been validated in many visual recognition tasks. To combine the two bag-of-features representations based on SIFT and HoG–HoF, we train separate classifiers and simply average their probability predictions. From our evaluation, our scene models perform very similar to that in [5].

2) *Object-Level Action Atoms:* Besides the background scene, another useful contextual cue is the types of objects and their interactions in action videos. We, therefore, employ a state-of-the-art generic object detector [28], with models trained for locating *Person*, *Car*, and *Chair*. Since our aim is to learn human actions, *Person* is obviously the most important object of interest, while other objects such as *Car* and *Chair* are helpful for identifying actions involving object interactions (e.g., *person getting out of car* and *standing up*). Object detectors other than the three can be easily deployed in our framework without any significant modification.

The output of the detectors is a set of object bounding boxes over static video frames. Since temporal information is not considered during detection, the results are not always consistent across nearby frames—there are many false alarms and miss-detections. To alleviate the effect of this issue, we track the detected objects based on their spatial locations and bounding box sizes (bounding box overlap must exceed 40%), and discard the isolated detections without any tracked association within a temporal window. The window size is empirically set as 15 frames (0.5 s).

Based on the detection, here we define ten object-level action atoms, including single object, multiple objects, as well as object interactions (with varying spatial distances along time). Fig. 3 gives an exemplar for each of the atoms. We expect that knowing the presence of these atoms, though noisy, will be helpful for recognizing human actions (e.g.,

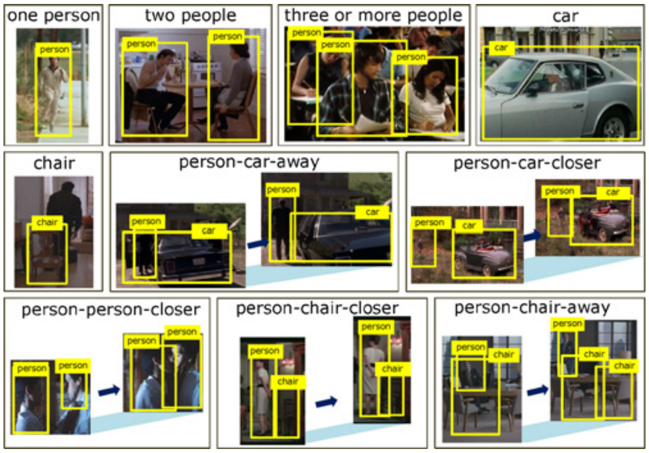


Fig. 3. Object-level action atoms. The first five indicate object presence and/or count, and the rest describes object interactions.

*person-person-closer* for action *kissing*). To characterize the object interactions, we compute average spatial distance between different types of detected objects (or among all detected people for *person-person-closer*). The average distance is compared to that in the next successive frame in order to evaluate the object proximity changes. Since the detection of an atom from a single frame can be noisy, we evaluate the presence of each atom based on overall statistics on the entire frame sequence. Specifically, the probability that an atom occurs in a video is estimated based on the percentage of frames where the atom is detected (e.g., the percentage of the frames containing two detected people for atom *two people*).

### C. Estimating Action-Scene-Object Relationship

In this section, we introduce an algorithm to estimate the action-scene-object dependency from limited samples. Inspired by the popularity of semi-supervised classification and recent works on semi-supervised ensemble learning [29], [30], we consider both labeled and unlabeled data for modeling the action context. Different from the LGC method [27], our goal here is to reveal how scene and object contexts correlate with human actions, not to propagate labels for classification/retrieval. We describe the details of this algorithm below.

Let $\mathcal{X}$ be a training dataset of $n$ samples, both labeled and unlabeled, and $F$ be an $n \times m$ prediction matrix of the contextual cues (i.e., scene classes or object atoms) over the $n$ training samples. $m$ is the number of contextual cues. Given an action, denote $\mathcal{P}$ as a positive training sample set (the given examples), and $\mathcal{N}$ as a set of negative samples $[(\mathcal{P} \cup \mathcal{N}) \subset \mathcal{X}]$. $\mathcal{N}$ can be automatically collected based on the method introduced earlier. In the scenario considered in this paper, the cardinality of $\mathcal{P}$, i.e., the number of positive labeled samples, should be much smaller than $n$.

Our aim is to derive a set of coefficients based on which the linear combination of the columns in $F$ (contextual cues) satisfies two conditions. First, it can well distinguish samples from $\mathcal{P}$ and $\mathcal{N}$. Second, it should produce similar scores if two samples are close or on the same manifold in context feature space ($\mathbb{R}^m$). A linear model is applied here due to its better generalization capability especially when the amount of training data is limited [31]. More formally, we define an action inference function as follows:

$$r : \mathcal{X} \rightarrow \mathbb{R} \tag{1}$$

where $r(\mathbf{x})$, $\mathbf{x} \in \mathcal{X}$, is the context-based inference score (from the linear combination of contextual cues) for sample $\mathbf{x}$.

To satisfy the first condition, the inference scores of the positive samples should be higher than those of the negative ones, that is

$$r(\mathbf{x}_i) > r(\mathbf{x}_j) \Longleftrightarrow i \in \mathcal{P}, j \in \mathcal{N}. \tag{2}$$

Based on the above analysis, we assume $r$ can be obtained by linearly combining the columns of $F$, that is

$$[r(\mathbf{x}_1), \dots, r(\mathbf{x}_n)]^\top = F\mathbf{c} \tag{3}$$

where $\mathbf{c}$ denotes the coefficient vector. Note that $r(\mathbf{x}_i) = \mathbf{c}^\top \mathbf{f}_i$, where $\mathbf{f}_i^\top$ is the $i$th row vector of $F$, i.e., the predictions of the contextual cues on sample $\mathbf{x}_i$. Now the aim is to obtain a $\mathbf{c}$ that best satisfies the inequality defined in (2), which can be rewritten as

$$\mathbf{c}^\top \mathbf{f}_i - \mathbf{c}^\top \mathbf{f}_j \geq 1, i \in \mathcal{P}, j \in \mathcal{N}. \tag{4}$$

In other words, we want the inference score of a positive sample to be higher than that of any negative sample by a hard margin 1. Note that the margin can be arbitrary positive number, as it only corresponds to a scaling factor of $\mathbf{c}$. It is possible that no $\mathbf{c}$ can meet all the constraints in (4). We, therefore, add a non-negative slack variable to each constraint as follows:

$$\begin{aligned} &\mathbf{c}^\top \mathbf{f}_i - \mathbf{c}^\top \mathbf{f}_j \geq 1 - \xi_{ij} \\ &\xi_{ij} \geq 0, i \in \mathcal{P}, j \in \mathcal{N}. \end{aligned} \tag{5}$$

On the contrary, if two samples are close or on the same manifold structure in the context feature space, it is expected that their inference scores are similar. This condition is generally referred to as the smoothness property, which is popularly used in many semi-supervised learning approaches. It can be measured by

$$\mathcal{S}(r) \triangleq \frac{1}{2} \sum_{i,j} w_{ij} (r(\mathbf{x}_i) - r(\mathbf{x}_j))^2 \tag{6}$$

where $w_{ij}$ is the similarity between any two samples (both labeled and unlabeled) $\mathbf{x}_i$ and $\mathbf{x}_j$. In our case, we take

$$w_{ij} = \exp^{-\Omega \|\mathbf{f}_i - \mathbf{f}_j\|^2} \tag{7}$$

where $\Omega > 0$ is a scaling factor. Clearly, the smaller of $\mathcal{S}(r)$, the smoother of the inference function $r$.

Plugging (3) into (6), $\mathcal{S}(r)$ can be expressed as

$$\mathcal{S}(r) = \frac{1}{2} \mathbf{c}^\top L \mathbf{c} \tag{8}$$

where $L = \sum_{i,j} w_{ij} (\mathbf{f}_i - \mathbf{f}_j)(\mathbf{f}_i - \mathbf{f}_j)^\top$. Taking a maximum margin approach, we now formulate the entire problem as the optimization process as follows:

$$\begin{aligned}
&\min_{\mathbf{c},\xi_{ij}} \frac{1}{2}\mathbf{c}^\top\mathbf{c} + \alpha\sum_{i,j}\xi_{ij} + \frac{\beta}{2}\mathbf{c}^\top L\mathbf{c} \\
&\text{s.t.}\mathbf{c}^\top\mathbf{f}_i - \mathbf{c}^\top\mathbf{f}_j \geq 1 - \xi_{ij}, \\
&\xi_{ij} \geq 0, i \in \mathcal{P}, j \in \mathcal{N}
\end{aligned} \tag{9}$$

where the first term is the reciprocal of the maximum margin, the second term is the fitting error of separating the positive and negative samples, and the last term is the smoothness constraint. $\alpha$ and $\beta$ are two nonnegative parameters that trade off the margin, the fitting error, and the smoothness. This is a strictly convex quadratic function in $\mathbf{c}$ and $\xi_{ij}$s, and all the constraints are linear. Therefore, it can be optimally solved by quadratic programming. Since the number of labeled samples is small in our problem, the optimal coefficients $\tilde{\mathbf{c}}$ can be efficiently computed.

Up to this point, we can determine the relationship of an arbitrary action to the scene and object contexts. It is worthwhile to note a fundamental difference between this algorithm and the linear SVMs [32]—a smoothness term is added here to utilize the large volume of readily available unlabeled data, which is important as the smoothness property in semi-supervised learning has been known to be suitable for learning from limited labeled samples.

### D. Incorporating Multiple Contextual Cues

We now turn to the problem of how to use the learned coefficients $\tilde{\mathbf{c}}$ for action discovery from new video data. Since our aim in the previous section was to optimize the coefficients of linearly fusing multiple contextual cues, given an action $a$, linear combination based on its optimal coefficients $\tilde{\mathbf{c}}_a$ becomes a natural solution as follows:

$$\begin{aligned}
\tilde{q}_a(\mathbf{x}) &= q_a(\mathbf{x}) + \lambda r(\mathbf{x}) \\
&= q_a(\mathbf{x}) + \lambda\tilde{\mathbf{c}}_a^\top\mathbf{f}_\mathbf{x}
\end{aligned} \tag{10}$$

where $\mathbf{x}$ is a test sample and $\mathbf{f}_\mathbf{x}$ is the prediction scores of the contextual cues on $\mathbf{x}$; $q_a(\mathbf{x})$ is baseline action prediction score based on direct comparison from the raw visual features (e.g., HoG–HoF); $\tilde{q}_a(\mathbf{x})$ is the refined prediction after incorporating the contextual cues; $\lambda$ is a context weight parameter. The baseline prediction $q_a(\mathbf{x})$ can be obtained via either supervised learning or semi-supervised learning, which will be elaborated in the experiments.

## IV. Experiments

In this section, we test our approach on the challenging Hollywood2 dataset [5], where the training and test sets contain 823 and 884 movie video clips, respectively (about 500k frames in total). Scene models are trained on a separate training set in Hollywood2, which is different from the action training/test data. We report performance over 12 actions defined in [5] with varying (small) amounts of action examples. Note that our approach is designed to handle arbitrary human actions with just a few examples. We experiment with the 12 actions since ground-truth action annotation is needed for performance evaluation. Results are measured using average precision over the entire test set, and mean average precision (mAP) is used to aggregate performance of multiple actions.

In the following, we first evaluate our algorithm on estimating action-scene-object relationship and compare it with existing alternatives. We then show the combination of context with baseline predictions using the raw visual features.

### A. Retrieval Performance by Contextual Cues

The goal of this experiment is to show how well the contextual cues alone can perform for action recognition. To this end, we set the baseline prediction $q_a(\mathbf{x})$ in (10) as 0. Given a few positive samples, we first automatically collect an equivalent number of negative samples using the method introduced in Section III-A. Throughout the experiments, the two parameters $\alpha$ and $\beta$ in (9) are both set as 10. Our evaluations indicate that the results are not sensitive to these parameters as long as relatively large values are used, to ensure that the learned coefficients ($\tilde{\mathbf{c}}$) fit well to the labeled data and produce smooth action inference scores [$r(\mathbf{x})$] in the context feature space.

Fig. 4(a) shows the mAP performances with different numbers of examples per action. For each method and each example number, we plot mean performance over ten runs with randomly selected examples. We see that both scene and object contexts perform significantly better than the prior (chance) even with just a single label. This confirms that the contextual cues are useful for human action retrieval. In general, scene context shows higher performance than object context. This is probably due to the fact that the object detections are noisy. Table I further shows the per-action performance when ten examples are used for each action. All the actions substantially benefit from both types of context. Only using the scene context, we obtain fairly good performance for several actions, such as *person driving car* and *running*. From investigating the learned coefficients, we observe close relatedness of scene classes *Car Interior* and *Road* to the actions *person driving car* and *running*, respectively. On the contrary, the object context works better than scene context for a few actions, e.g., *people shaking hands*, which tightly correlates to *person-person-closer* and *three-or-more-people* as indicated by the learned coefficients.

1) *Comparison to the State of the Arts:* We compare our approach with the two strategies for context-based action recognition proposed in [5]: 1) SVM learning, and 2) movie script-mining. The first strategy treats predictions of the contextual cues ($\mathbf{f}_\mathbf{x}$) as a video feature vector, on which an SVM classifier can be trained for each action. Since the SVM learning is fully supervised, this strategy requires a large number of labeled training samples in order to obtain satisfactory performance. In contrast, the script-mining method estimates action context based on the co-occurrence of action and scene names in textual movie scripts, which does not require labeled data. For this method, we use the conditional probabilities $p(action|scene)$ from [5][1] to replace the coefficients in $\tilde{\mathbf{c}}$. Since only action-scene relationship is

[1] Available at http://pascal.inrialpes.fr/hollywood2.

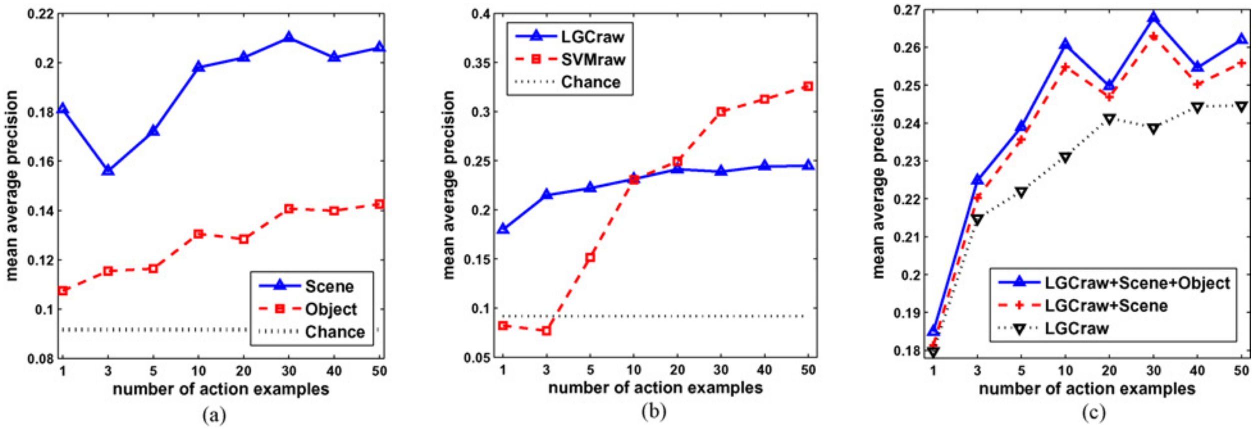


Fig. 4. (a) Action retrieval performance using contextual cues alone. With only ten examples per action, our context modeling approach produces mAPs of 19.8% (scene) and 13.5% (object), compared to the mean prior of 9.2% (chance). (b) Comparison of baseline retrieval performance using raw visual features from semi-supervised learning (LGCraw) and support vector machines (SVMraw). (c) Combining our context-based retrieval with the LGCraw baseline.

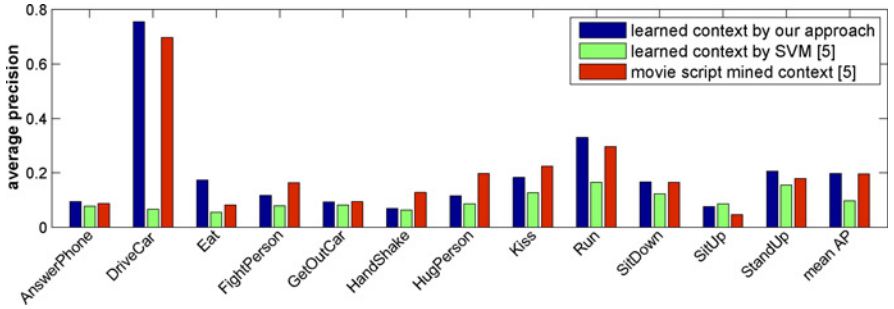


Fig. 5. Comparison of different *context-based action retrieval* methods. Both our approach (blue bars) and the SVM-based method (green bars) use ten examples per action. Our approach performs much better than the fully supervised SVM learning. More interestingly, with just ten labels, it already shows slightly better performance than the movie script-mining method from [5] (red bars), which is a domain-specific method and not suitable for other data domains. See Section IV-A1 for more explanations.

available from [5], we use the scene context alone in this experiment.

Fig. 5 compares the results of our approach and the other two strategies. Our approach significantly outperforms the SVM learning method. This again confirms that the proposed algorithm incorporated ideas from semi-supervised learning is more suitable to model action context when the number of examples is small. Using just ten labels per action, we obtain slightly better performance (mAP=19.8%) than the script-mining method (mAP=19.6%)—this shows an important merit of our approach, since the latter is designed particularly for movie videos and, therefore, the mined contexts may be too domain-specific to be applied for data from other domains (e.g., consumer videos on the Internet). The ability of rapid context modeling with few examples enables the possibility of learning arbitrary actions from other data genres, which is of broad interest in many practical applications.

Our context modeling algorithm is very efficient—on a regular personal computer with a 3 GHz central processing unit, it requires 10–20 s to compute the optimal coefficients for each action, depending on the number of examples used.

### B. Combining Context with Baseline Visual Matching

We now turn to the problem of combining context-based retrieval with the baseline predictions $q_a(\mathbf{x})$ from direct comparison of the raw visual features. We consider two ways for computing $q_a(\mathbf{x})$.

TABLE I
CONTEXT-ONLY PERFORMANCE FOR EACH OF THE EVALUATED ACTIONS (TEN EXAMPLES)

| | Scene | Object | *Chance* |
|---|---|---|---|
| *AnswerPhone* | 0.094 | 0.077 | *0.072* |
| *DriveCar* | 0.755 | 0.274 | *0.115* |
| *Eat* | 0.173 | 0.073 | *0.037* |
| *FightPerson* | 0.117 | 0.114 | *0.079* |
| *GetOutCar* | 0.093 | 0.113 | *0.065* |
| *HandShake* | 0.069 | 0.124 | *0.051* |
| *HugPerson* | 0.115 | 0.078 | *0.075* |
| *Kiss* | 0.183 | 0.121 | *0.117* |
| *Run* | 0.330 | 0.175 | *0.160* |
| *SitDown* | 0.166 | 0.179 | *0.122* |
| *SitUp* | 0.076 | 0.059 | *0.042* |
| *StandUp* | 0.206 | 0.179 | *0.165* |

1) *SVMraw*: SVM learning based on the raw visual features using the $\chi^2$ Gaussian kernel, as has been used in [3], [5] and [21].
2) *LGCraw*: Semi-supervised learning by the LGC method [27] which propagates labels to the unlabeled test data.

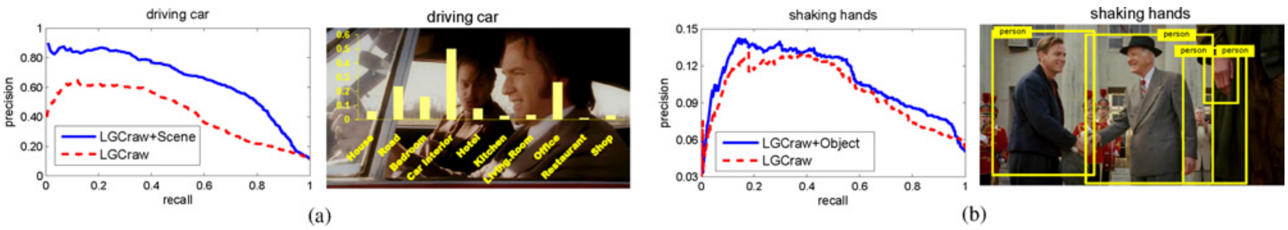


Fig. 6. Precision-recall curves and example frames of two actions for which the context cues significantly improve retrieval performance. (a) *Driving car*, where the scene context (e.g., *Car Interior*) is particularly helpful. (b) *Shaking hands*, which benefits more from the object context. Scene and object detection results are shown on the two example frames, respectively.

For LGCraw, we can reuse its result from the process of collecting negative samples (see Section III-A). We only use the HoG–HoF feature, since the 2-D SIFT features do not contribute much for action recognition as was also observed in [5].

We first compare the two baseline methods SVMraw and LGCraw with varying numbers of examples. Results are visualized in Fig. 4(b). As expected, LGCraw shows better performance when the number of examples is small. This can be explained by the fact that semi-supervised learning ensures the smoothness of prediction scores according to the manifold structure of sample distribution in the HoG–HoF feature space, where the large amount of unlabeled samples is utilized.

We also evaluate the performance gain when combining predictions from the contextual cues and the LGCraw baseline. For the context weight $\lambda$, we empirically set it to 0.1. Smaller $\lambda$ is preferred since context serves as auxiliary information for improving the baseline visual matching. Fig. 4(c) gives the results. We see the scene context contributes significantly at most of the example numbers. The object context, on the contrary, further provides less but consistent improvements. Combining context and the baseline raw feature matching, we obtain an mAP of 26% when only ten examples are used. We consider this as a good achievement, compared to the results (mAP 30–40%) reported in [5] and [21], where hundreds of training samples are used. When similar amount of examples are in use, we also observed better performance close to 40%, which is out of the scope of this paper, however. Again, note that [5] used movie script-mining to compute the action-scene relationship, while our approach is more generic and can be applied to any type of data.

Fig. 6 further shows precision–recall curves and visual exemplars of two actions that benefit from the use of the scene and object contexts, respectively.

## V. Conclusion

We have presented a framework for human action retrieval based on very few examples, which has tremendous potential in many practical applications. An algorithm based on the popular semi-supervised learning paradigm was introduced to model action-scene-object dependency from limited examples. Our experimental results showed that, with the learned dependency information, both object and background scene contexts are useful for action retrieval. Particularly, using just a few examples per action, our approach outperformed an existing domain-specific method based on movie script mining. Further, coupling the context-based action prediction with baseline learning from raw visual features, substantial performance gains are observed.

Currently, the performance gain from object context is smaller than that from scene context due to noisy detection. Therefore, an interesting future work is to evaluate how detection accuracy of the contextual cues, particularly the object context, affects the action retrieval performance. This may provide useful insights in understanding "how good is object detection good enough for significantly helping real-world video retrieval applications."

## References

[1] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. Int. Conf. Patt. Recog.*, vol. 3. 2004, pp. 32–36.

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. Int. Conf. Comput. Vision*, vol. 2. 2005, pp. 1395–1402.

[3] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. Conf. Comput. Vision Patt. Recog.*, 2008, pp. 1–8.

[4] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos 'in the wild,'" in *Proc. Conf. Comput. Vision Patt. Recog.*, 2009, pp. 1996–2003.

[5] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. Conf. Comput. Vision Patt. Recog.*, 2009, pp. 2929–2936.

[6] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. Conf. Comput. Vision Patt. Recog.*, 2006, pp. 2169–2178.

[7] I. Laptev and T. Lindeberg, "Space-time interest point," in *Proc. Int. Conf. Comput. Vision*, vol. 1. 2003, pp. 432–439.

[8] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. Workshop Vis. Surveillance Performance Eval. Track. Surveillance*, 2005, pp. 65–72.

[9] A. Klaser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3-D-gradients," in *Proc. British Mach. Vision Conf.*, 2008, pp. 995–1004.

[10] J. Sun, X. Wu, S. Yan, L. F. Cheong, T. S. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition," in *Proc. Conf. Comput. Vision Patt. Recog.*, 2009, pp. 2004–2011.

[11] H. J. Seo and P. Milanfar, "Detection of human actions from a single example," in *Proc. Int. Conf. Comput. Vision*, 2009, pp. 1965–1970.

[12] D. Tran and A. Sorokin, "Human activity recognition with metric learning," in *Proc. Eur. Conf. Comput. Vision*, 2008, pp. 548–561.

[13] C. Wu and H. Aghajan, "Using context with statistical relational models: Object recognition from observing user activity in home environment," in *Proc. Workshop Use Context Vision Process.*, 2009, pp. 22–27.

[14] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[15] A. Torralba, "Contextual priming for object detection," *Int. J. Comput. Vision*, vol. 53, no. 2, pp. 169–191, 2003.

[16] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *Proc. Int. Conf. Comput. Vision*, 2007, pp. 1–8.

[17] B. Russell, A. Torralba, C. Liu, R. Fergus, and W. Freeman, "Object recognition by scene alignment," in *Proc. Neural Inform. Process. Syst.*, 2007, pp. 1241–1248.

[18] A. Gupta and L. Davis, "Objects in action: An approach for combining action understanding and object perception," in *Proc. Conf. Comput. Vision Patt. Recog.*, 2007, pp. 1–8.

[19] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. Rehg, "A scalable approach to action recognition based on object use," in *Proc. Int. Conf. Comput. Vision*, 2007, pp. 361–368.

[20] S. Tran and L. S. Davis, "Visual event modeling and recognition using Markov logic networks," in *Proc. Eur. Conf. Comput. Vision*, 2008, pp. 610–623.

[21] D. Han, L. Bo, and C. Sminchisescu, "Selection and context for action recognition," in *Proc. Int. Conf. Comput. Vision*, 2009, pp. 1933–1940.

[22] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *Proc. Conf. Comput. Vision Patt. Recog.*, 2010, pp. 17–24.

[23] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, Apr. 2006.

[24] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *Int. J. Comput. Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[25] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. British Mach. Vision Conf.*, 2009, pp. 127–137.

[26] Y. G. Jiang, C. W. Ngo, and J. Yang, "Toward optimal bag-of-features for object categorization and semantic video retrieval," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2007, pp. 494–501.

[27] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Scholkopf, "Learning with local and global consistency," in *Proc. Neural Inform. Process. Syst.*, 2004, pp. 321–328.

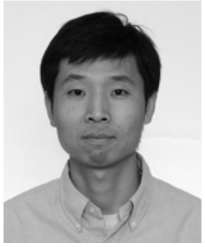
[28] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. Conf. Comput. Vision Patt. Recog.*, 2008, pp. 1–8.

[29] S. C. H. Hoi and R. Jin, "Semi-supervised ensemble ranking," in *Proc. AAAI Conf. Artif. Intell.*, 2008, pp. 634–639.

[30] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, Dec. 2006.

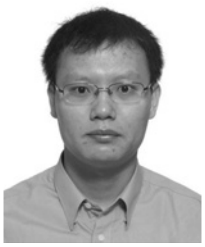
[31] D. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[32] T. Joachims, "Training linear SVMs in linear time," in *Proc. ACM SIGKDD Conf. Knowledge Discovery Data Mining*, 2006, pp. 217–226.



**Yu-Gang Jiang** received the Ph.D. degree in computer science from the City University of Hong Kong, Kowloon, Hong Kong, in 2009. During his Ph.D. dissertation research, he spent a year studying at Columbia University, New York, NY, as a Visiting Scholar.

He is currently a Post-Doctoral Research Scientist with the Department of Electrical Engineering, Columbia University. His current research interests include multimedia retrieval, computer vision, and visual/sound perception. He has authored more than 20 papers in these fields. He is an active participant of the Annual NIST TRECVID Evaluation and has designed a few top-performing video retrieval systems over the years.



**Zhenguo Li** received the B.S. and M.S. degrees from the Department of Mathematics, Peking University, Beijing, China, in 2002 and 2005, respectively, and the Ph.D. degree from the Department of Information Engineering, Chinese University of Hong Kong, Shatin, Hong Kong, in 2008.

He is currently a Post-Doctoral Research Scientist with the Department of Electrical Engineering, Columbia University, New York, NY. His current research interests include computer vision and machine learning.

**Shih-Fu Chang** (S'89–M'90–SM'01–F'04) is a Professor of electrical engineering and the Director of the Digital Video and Multimedia Laboratory, Columbia University, New York, NY. From 2007 to 2010, he was the Chair of the Department of Electrical Engineering, Columbia University. He has made significant contributions to multimedia searches, visual communication, media forensics, and international standards for multimedia. He was with different advising/consulting capacities for IBM, Microsoft, Kodak, PictureTel, and several other institutions.

He has been the recipient of several awards, including the IEEE Kiyo Tomiyasu Award, the Navy ONR Young Investigator Award, the IBM Faculty Award, the ACM Recognition of Service Award, and the NSF CAREER Award. He and his students have received many Best Paper and Best Student Paper Awards from IEEE, ACM, and SPIE. From 2006 to 2008, he was the Editor-in-Chief for the IEEE SIGNAL PROCESSING MAGAZINE.